An Introduction to Efficiency and Productivity Analysis

Second Edition

Timothy J. Coelli, D.S. Prasada Rao
Christopher J. O'Donnell
and George E. Battese

Springer

AN INTRODUCTION TO EFFICIENCY AND PRODUCTIVITY ANALYSIS

Second Edition

AN INTRODUCTION TO EFFICIENCY AND PRODUCTIVITY ANALYSIS

Second Edition

by

Timothy J. Coelli
D.S. Prasada Rao
Christopher J. O'Donnell
George E. Battese

Springer

Tim Coelli
University of Queensland
Australia

D.S. Prasada Rao
University of Queensland
Australia

Christopher J. O'Donnell
University of Queensland
Australia

George E. Battese
University of Queensland
Australia

Library of Congress Cataloging-in-Publication Data

An introduction to efficiency and productivity analysis / by Timothy Coelli ... [et al].—
2nd ed.
p. cm.
Rev. ed. of: An introduction to efficiency and productivity analysis / by Tim Coelli. c1998.
Includes bibliographical references and index.
ISBN-10: 0-387-24265-1 ISBN-13: 978-0387-24265-1
ISBN-10: 0-387-24266-X ISBN-13: 978-0387-24266-8 (softcover)
e-ISBN-10: 0-387-25895-7
1. Production (Economic theory) 2. Production functions (Economic theory) 3. Industrial productivity. I. Coelli, Tim. II. Coelli, Tim. Introduction to efficiency and productivity analysis.

HB241.C64 2005
338'.06—dc22

2005042642

Copyright © 2005 by Springer Science+Business Media, Inc.
All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science + Business Media, Inc., 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now know or hereafter developed is forbidden.
The use in this publication of trade names, trademarks, service marks and similar terms, even if the are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1 SPIN 11053217

springeronline.com

To

Michelle, Visala, Adrienne and Marilyn

TABLE OF CONTENTS

List of Figures *page xi*
List of Tables *xiii*
Preface *xv*

1. INTRODUCTION 1
   1.1 Introduction 1
   1.2 Some Informal Definitions 2
   1.3 Overview of Methods 6
   1.4 Outline of Chapters 7
   1.5 What is Your Economics Background? 9

2. REVIEW OF PRODUCTION ECONOMICS 11
   2.1 Introduction 11
   2.2 Production Functions 12
   2.3 Transformation Functions 20
   2.4 Cost Functions 21
   2.5 Revenue Functions 31
   2.6 Profit Functions 32
   2.7 Conclusions 40

3. PRODUCTIVITY AND EFFICIENCY MEASUREMENT CONCEPTS 41
   3.1 Introduction 41
   3.2 Set Theoretic Representation of a Production Technology 42
   3.3 Output and Input Distance Functions 47
   3.4 Efficiency Measurement using Distance, Cost and Revenue Functions 51
   3.5 Measuring Productivity and Productivity Change 61
   3.6 Conclusions 81

4. INDEX NUMBERS AND PRODUCTIVITY MEASUREMENT 85
   4.1 Introduction 85
   4.2 Conceptual Framework and Notation 86
   4.3 Formulae for Price Index Numbers 88
   4.4 Quantity Index Numbers 90
   4.5 Properties of Index Numbers: The Test Approach 95
   4.6 The Economic-Theoretic Approach 98
   4.7 A Simple Numerical Example 113
   4.8 Transitivity in Multilateral Comparisons 116
   4.9 TFP Change Measurement Using Index Numbers 118
   4.10 Empirical Application: Australian National Railways 127
   4.11 Conclusions 131

5. DATA AND MEASUREMENT ISSUES 133
5.1 Introduction 133
5.2 Outputs 135
5.3 Inputs 141
5.4 Prices 153
5.5 Comparisons over time 154
5.6 Output aggregates for sectoral and economy-wide comparisons 156
5.7 Cross-country comparisons of productivity 157
5.8 Data editing and errors 159
5.9 Conclusions 160

6. DATA ENVELOPMENT ANALYSIS 161
6.1 Introduction 161
6.2 The Constant Returns to Scale DEA Model 162
6.3 The Variable Returns to Scale Model and Scale Efficiencies 172
6.4 Input and Output Orientations 180
6.5 Conclusions 181

7. ADDITIONAL TOPICS ON DATA ENVELOPMENT ANALYSIS 183
7.1 Introduction 183
7.2 Price Information and Allocative Efficiency 183
7.3 Non-Discretionary Variables 188
7.4 Adjusting for the Environment 190
7.5 Input Congestion 195
7.6 Treatment of Slacks 198
7.7 Additional Methods 199
7.8 Empirical Application: Australian Universities 203
7.9 Conclusions 206

8. ECONOMETRIC ESTIMATION OF PRODUCTION TECHNOLOGIES 209
8.1 Introduction 209
8.2 Production, Cost and Profit Functions 210
8.3 Single Equation Estimation 214
8.4 Imposing Equality Constraints 220
8.5 Hypothesis Testing 223
8.6 Systems Estimation 225
8.7 Inequality Constraints 227
8.8 The Bayesian Approach 231
8.9 Simulation Methods 234
8.10 Conclusion 239

9. STOCHASTIC FRONTIER ANALYSIS 241
9.1 Introduction 241
9.2 The Stochastic Production Frontier 242
9.3 Estimating the Parameters 245
9.4 Predicting Technical Efficiency 254
9.5 Hypothesis Testing 258
9.6 Conclusions 261

10. ADDITIONAL TOPICS ON STOCHASTIC FRONTIER ANALYSIS 263
10.1 Introduction 263
10.2 Distance Functions 264
10.3 Cost Frontiers 266
10.4 Decomposing Cost Efficiency 269
10.5 Scale Efficiency 272
10.6 Panel Data Models 275
10.7 Accounting for the Production Environment 281
10.8 The Bayesian Approach 284
10.9 Conclusions 288

11. THE CALCULATION AND DECOMPOSITION OF PRODUCTIVITY CHANGE USING FRONTIER METHODS 289
11.1 Introduction 289
11.2 The Malmquist TFP Index and Panel Data 291
11.3 Calculation using DEA Frontiers 294
11.4 Calculation using SFA Frontiers 300
11.5 An Empirical Application 302
11.6 Conclusions 310

12. CONCLUSIONS 311
12.1 Summary of Methods 311
12.2 Relative Merits of the Methods 312
12.3 Some Final Points 313

Appendix 1: Computer Software *317*
Appendix 2: Philippines Rice Data *325*
References *327*
Author Index *341*
Subject Index *345*

FIGURES

1.1	Production Frontiers and Technical Efficiency	4
1.2	Productivity, Technical Efficiency and Scale Economies	5
1.3	Technical Change Between Two Periods	6
2.1	Single-Input Production Function	14
2.2	Output Isoquants	15
2.3	A Family of Production Functions	15
2.4	Elasticities of Substitution	17
2.5	Short-Run Production Functions	21
2.6	Cost Minimisation	24
2.7	Long-Run and Short-Run Fixed, Variable and Total Costs	29
2.8	Profit Maximisation	35
2.9	LTR, LTC and Profit Maximisation	36
2.10	LMR, LMC and Profit Maximisation	38
3.1	Production Possibility Curve	45
3.2	The Production Possibility Curve and Revenue Maximisation	46
3.3	Technical Change and the Production Possibility Curve	46
3.4	Output Distance Function and Production Possibility Set	48
3.5	Input Distance Function and Input Requirement Set	50
3.6	Technical and Allocative Efficiencies	52
3.7	Input- and Output-Orientated Technical Efficiency Measures and Returns to Scale	55
3.8	Technical and Allocative Efficiencies from an Output Orientation	55
3.9	The Effect of Scale on Productivity	59
3.10	Scale Efficiency	61
3.11	Malmquist Productivity Indices	71
4.1	Revenue Maximisation	100
4.2	Output Price Index	101
4.3	Input Price Index	105
4.4	Indices of Output, Input and TFP for Australian National Railways	130
5.1	Age efficiency profiles under different assumptions	148
6.1	Efficiency Measurement and Input Slacks	165
6.2	CRS Input-Orientated DEA Example	167
6.3:	Scale Efficiency Measurement in DEA	174
6.4	VRS Input-Orientated DEA Example	175
6.5	Output-Orientated DEA	181
7.1	CRS Cost Efficiency DEA Example	187
7.2	Efficiency Measurement and Input Disposability (Congestion)	197
7.3	Super Efficiency	201

8.1 The Metropolis-Hastings Algorithm 238
9.1 The Stochastic Production Frontier 244
9.2 Half-Normal Distributions 247
9.3 Truncated-Normal Distributions 254
10.1 Functions for Time-Varying Efficiency Models 278
11.1 Malmquist DEA Example 296
11.2 Cumulative Percentage Change Measures of TEC, TC, SC and TFPC using SFA 305
11.3 Cumulative Percentage Change Measures of TEC, TC, SC and TFPC using DEA 307
11.4 Cumulative TFP Change using DEA, SFA and PIN 308

TABLES

4.1	Data for Billy's Bus Company	113
4.2a	SHAZAM Instructions for Output Price and Quantity Indices	114
4.2b	SHAZAM Output for Output Price and Quantity Indices	115
4.3a	Listing of Data file, EX1.DTA	124
4.3b	Listing of Instruction File, EX1.INS	125
4.3c	Listing of Output File, EX1.OUT	125
4.4a	Listing of Instruction File, EX2.INS	126
4.4b	Listing of Output File, EX2.OUT	126
4.5	Output Data for the Australian National Railways Example	128
4.6	Non-capital Input Data for the Australian National Railways Example	128
4.7	Capital Input Data for the Australian National Railways Example	129
4.8	Indices of Output, Input and TFP for Australian National Railways	130
6.1	Example Data for CRS DEA Example	165
6.2	CRS Input-Orientated DEA Results	167
6.3a	Listing of Data File, EG1-DTA.TXT	168
6.3b	Listing of Instruction File, EG1-INS.TXT	169
6.3c	Listing of Output File, EG1-OUT.TXT	169
6.4	Example Data for VRS DEA	175
6.5	VRS Input-Orientated DEA Results	176
6.6a	Listing of Data File, EG2-DTA.TXT	176
6.6b	Listing of Instruction File, EG2-INS.TXT	177
6.6c	Listing of Output File, EG2-OUT.TXT	177
7.1	CRS Cost Efficiency DEA Results	187
7.2a	Listing of Data File, EG3-DTA.TXT	188
7.2b	Listing of Instruction File, EG3-INS.TXT	188
7.2c	Listing of Output File, EG3-OUT.TXT	189
7.3	DEA Results for the Australian Universities Study	205
8.1	Some Common Functional Forms	211
8.2	OLS Estimation of a Translog Production Function	216
8.3	NLS Estimation of a CES Production Function	219
8.4	Constant Returns to Scale Translog Production Function	222
8.5	Systems Estimation of a Translog Cost Function	228
8.6	Imposing Global Concavity on a Translog Cost Function	230
8.7	Bayesian Estimation of a Translog Production Function	236
8.8	Monotonicity-Constrained Translog Production Function	238
9.1	Estimating a Half-Normal Frontier Using SHAZAM	248
9.2	The FRONTIER Instruction File, CHAP9_2.INS	249
9.3	The FRONTIER Data File, CHAP9.TXT	249

9.4	The FRONTIER Output File For The Half-Normal Frontier	250
9.5	Estimating a Half-Normal Frontier Using LIMDEP	251
9.6	Estimating an Exponential Frontier Using LIMDEP	253
9.7	Predicting Firm-Specific Technical Efficiency Using SHAZAM	256
9.8	Predicting Industry Technical Efficiency Using SHAZAM	257
9.9	Estimating a Truncated-Normal Frontier Using FRONTIER	260
10.1	Estimating a Translog Cost Frontier Using SHAZAM	268
10.2	Decomposing Cost Efficiency Using SHAZAM	273
10.3	Truncated-Normal Frontier With Time-Invariant Inefficiency Effects	277
10.4	Truncated-Normal Frontier With Time-Varying Inefficiency Effects	280
10.5	Bayesian Estimation of an Exponential Frontier	287
11.1	Example Data for Malmquist DEA	296
11.2a	Listing of Data File, EG4-DTA.TXT	297
11.2b	Listing of Instruction File, EG4-INS.TXT	297
11.2c	Listing of Output File, EG4-OUT.TXT	298
11.3	Maximum-Likelihood Estimates of the Stochastic Frontier Model	303
11.4	Cumulative Percentage Change Measures of TEC, TC, SC and TFPC using SFA	304
11.5	Cumulative Percentage Change Measures of TEC, TC, SC and TFPC using DEA	307
11.6	Sample Average Input Shares	309
12.1	Summary of the Properties of the Four Principal Methods	312

PREFACE

The second edition of this book has been written for the same audience as the first edition. It is designed to be a "first port of call" for people wishing to study efficiency and productivity analysis. The book provides an accessible introduction to the four principal methods involved: econometric estimation of average response models; index numbers; data envelopment analysis (DEA); and stochastic frontier analysis (SFA). For each method, we provide a detailed introduction to the basic concepts, give some simple numerical examples, discuss some of the more important extensions to the basic methods, and provide references for further reading. In addition, we provide a number of detailed empirical applications using real-world data.

The book can be used as a textbook or as a reference text. As a textbook, it probably contains too much material to cover in a single semester, so most instructors will want to design a course around a subset of chapters. For example, Chapter 2 is devoted to a review of production economics and could probably be skipped in a course for graduate economics majors. However, it should prove useful to undergraduate students and those doing a major in another field, such as business management or health studies.

There have been several excellent books written on performance measurement in recent years, including Färe, Grosskopf and Lovell (1985, 1994), Fried, Lovell and Schmidt (1993), Charnes et al (1995), Färe, Grosskopf and Russell (1998) and Kumbhakar and Lovell (2000). The present book is not designed to compete with these advanced-level books, but to provide a lower-level bridge to the material contained within them, as well as to many other books and journal articles written on this topic.

We believe this second edition remains a unique book in this field insofar as:

1. it is an introductory text;
2. it contains detailed discussion and comparison of the four principal methods for efficiency and productivity analysis; and

3. it provides detailed advice on computer programs that can be used to implement these methods. The book contains computer instructions and output listings for the SHAZAM, LIMDEP, TFPIP, DEAP and FRONTIER computer programs. More extensive listings of data and computer instruction files are available on the book website (www.uq.edu.au/economics/cepa/crob2005).

The first edition of this book was published in 1998. It grew out of a set of notes that were written for a series of short courses that the Centre for Efficiency and Productivity Analysis (CEPA) had designed for a number of government agencies in Australia in the mid 1990's. The success of the first edition was largely due to its focus on the provision of information for practitioners (rather than academic theorists), and also due to the valuable feedback and suggestions provided by those people who attended these early short courses.

In the subsequent years we have continued to present CEPA short courses to people in business and government, using the first edition as a set of course notes. However, in recent years we have noted that we have been supplying increasing quantities of "extra materials" at these courses, reflecting the number of significant advances that have occurred in this field since 1998. Hence, when the publisher approached us to write a second edition, we were keen to take the opportunity to update the book with this new material. We also took the opportunity to freshen some of the original material to reflect our maturing understanding of various topics, and to incorporate some of the excellent suggestions provided by many readers and short course participants over the past seven years.

Readers familiar with the first edition will notice a number of changes in this second edition. Structurally, the material included in various chapters has been re-organised to provide a more logical ordering of economic theory and empirical methods. A number of new empirical examples have also been provided. Separate chapters have now been devoted to data measurement issues (Chapter 5) and the econometric estimation of average response functions (Chapter 8).

Many other changes and additions have also been incorporated. For example, the parametric methods section has been updated to cover confidence intervals; testing and imposing regularity conditions; and Bayesian methods. The DEA section has been updated to cover weights restrictions; super efficiency; bootstrapping; short-run cost minimisation; and profit maximisation. Furthermore, the productivity growth section has been updated to cover the issues of shadow prices and scale effects.

We wish to thank the many people whose comments, feedback and discussions have contributed to improving our understanding of the material within this book. In particular we wish to thank our recent CEPA visitors: Knox Lovell, Bert Balk, Erwin Diewert, Rolf Färe and Shawna Grosskopf. Rolf and Shawna were visiting

during the final few weeks of writing, and were very generous with their time, reading a number of draft chapters and providing valuable comments.

Finally, we hope that you, the readers, continue to find this book useful in your studies and research, and we look forward to receiving your comments and feedback on this second edition.

Timothy J. Coelli
D.S. Prasada Rao
Christopher J. O'Donnell
George E. Battese

Centre for Efficiency and
Productivity Analysis
University of Queensland
Brisbane, Australia.

1. INTRODUCTION

1.1 Introduction

This book is concerned with measuring the performance of firms, which convert inputs into outputs. An example of a firm is a shirt factory that uses materials, labour and capital (inputs) to produce shirts (output). The performance of this factory can be defined in many ways. A natural measure of performance is a productivity ratio: the ratio of outputs to inputs, where larger values of this ratio are associated with better performance. Performance is a relative concept. For example, the performance of the factory in 2004 could be measured relative to its 2003 performance or it could be measured relative to the performance of another factory in 2004, etc.

The methods of performance measurement that are discussed in this book can be applied to a variety of "firms".[1] They can be applied to private sector firms producing goods, such as the factory discussed above, or to service industries, such as travel agencies or restaurants. The methods may also be used by a particular firm to analyse the relative performance of units within the firm (e.g., bank branches or chains of fast food outlets or retail stores). Performance measurement can also be applied to non-profit organisations, such as schools or hospitals.

[1]In some of the literature on productivity and efficiency analysis the rather ungainly term "decision making unit" (DMU) is used to describe a productive entity in instances when the term "firm" may not be entirely appropriate. For example, when comparing the performance of power plants in a multi-plant utility, or when comparing bank branches in a large banking organisation, the units under consideration are really *parts* of a firm rather than firms themselves. In this book we have decided to use the term "firm" to describe any type of decision making unit, and ask that readers keep this more general definition in mind as they read the remainder of this book.

All of the above examples involve micro-level data. The methods we consider can also be used for making performance comparisons at higher levels of aggregation. For example, one may wish to compare the performance of an industry over time or across geographical regions (e.g., shires, counties, cities, states, countries, etc.).

We discuss the use and the relative merits of a number of different performance measurement methods in this book. These methods differ according to the type of measures they produce; the data they require; and the assumptions they make regarding the structure of the production technology and the economic behaviour of decision makers. Some methods only require data on quantities of inputs and outputs while other methods also require price data and various behavioural assumptions, such as cost minimisation, profit maximisation, etc.

But before we discuss these methods any further, it is necessary for us to provide some informal definitions of a few terms. These definitions are not very precise, but they are sufficient to provide readers, new to this field, some insight into the sea of jargon in which we swim. Following this we provide an outline of the contents of the book and a brief summary of the principal performance measurement methods that we consider.

1.2 Some Informal Definitions

In this section we provide a few informal definitions of some of the terms that are frequently used in this book. More precise definitions will be provided later in the book. The terms are:

- productivity;
- technical efficiency;
- allocative efficiency;
- technical change;
- scale economies;
- total factor productivity (TFP);
- production frontier; and
- feasible production set.

We begin by defining the **productivity** of a firm as the ratio of the output(s) that it produces to the input(s) that it uses.

$$\text{productivity} = \text{outputs/inputs} \tag{1.1}$$

When the production process involves a single input and a single output, this calculation is a trivial matter. However, when there is more than one input (which is

often the case) then a method for aggregating these inputs into a single index of inputs must be used to obtain a ratio measure of productivity.[2] In this book, we discuss some of the methods that are used to aggregate inputs (and/or outputs) for the construction of productivity measures.

When we refer to productivity, we are referring to **total factor productivity**, which is a productivity measure involving all factors of production.[3] Other traditional measures of productivity, such as labour productivity in a factory, fuel productivity in power stations, and land productivity (yield) in farming, are often called *partial* measures of productivity. These partial productivity measures can provide a misleading indication of overall productivity when considered in isolation.

The terms, **productivity** and **efficiency,** have been used frequently in the media over the last ten years by a variety of commentators. They are often used interchangeably, but this is unfortunate because they are not precisely the same things. To illustrate the distinction between the terms, it is useful to consider a simple production process in which a single input (x) is used to produce a single output (y). The line 0F′ in Figure 1.1 represents a **production frontier** that may be used to define the relationship between the input and the output. The production frontier represents the maximum output attainable from each input level. Hence it reflects the current state of technology in the industry. More is stated about its properties in later sections. Firms in this industry operate either on that frontier, if they are **technically efficient**, or beneath the frontier if they are not technically efficient. Point A represents an inefficient point whereas points B and C represent efficient points. A firm operating at point A is inefficient because technically it could increase output to the level associated with the point B without requiring more input.[4]

We also use Figure 1.1 to illustrate the concept of a **feasible production set** which is the set of all input-output combinations that are feasible. This set consists of all points between the production frontier, OF', and the x-axis (inclusive of these bounds).[5] The points along the production frontier define the efficient subset of this feasible production set. The primary advantage of the *set representation* of a production technology is made clear when we discuss multi-input/multi-output production and the use of distance functions in later chapters.

---

[2]The same problem occurs with multiple outputs.

[3] It also includes all outputs in a multiple-output setting.

[4]Or alternatively, it could produce the same level of output using less input (i.e., produce at point C on the frontier).

[5] Note that this definition of the production set assumes free disposability of inputs and outputs. These issues will be discussed further in subsequent chapters.

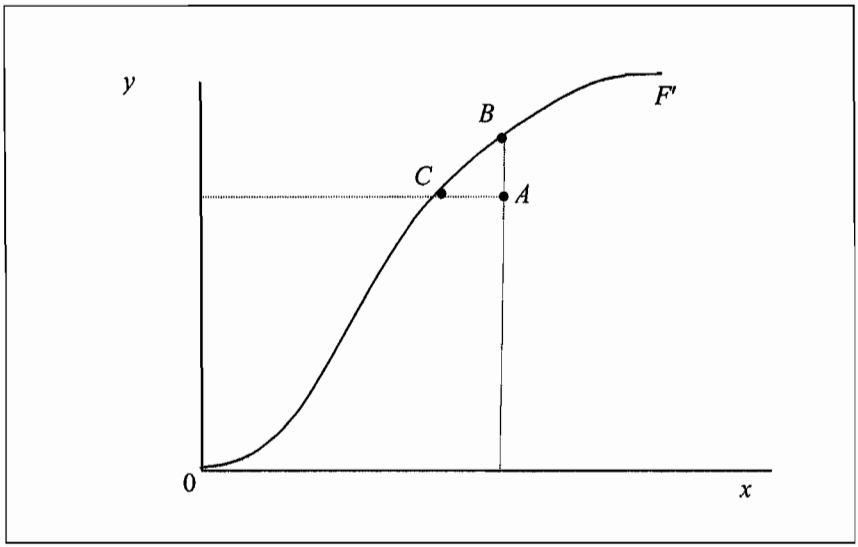

Figure 1.1 Production Frontiers and Technical Efficiency

To illustrate the distinction between technical efficiency and productivity we utilise Figure 1.2. In this figure, we use a ray through the origin to measure productivity at a particular data point. The slope of this ray is y/x and hence provides a measure of productivity. If the firm operating at point A were to move to the technically efficient point B, the slope of the ray would be greater, implying higher productivity at point B. However, by moving to the point C, the ray from the origin is at a tangent to the production frontier and hence defines the point of maximum possible productivity. This latter movement is an example of exploiting **scale economies**. The point C is the point of (technically) optimal scale. Operation at any other point on the production frontier results in lower productivity.

From this discussion, we conclude that a firm may be technically efficient but may still be able to improve its productivity by exploiting scale economies. Given that changing the scale of operations of a firm can often be difficult to achieve quickly, technical efficiency and productivity can in some cases be given short-run and long-run interpretations.

The discussion above does not include a time component. When one considers productivity comparisons through time, an additional source of productivity change, called **technical change,** is possible. This involves advances in technology that may be represented by an upward shift in the production frontier. This is depicted in Figure 1.3 by the movement of the production frontier from $0F_0'$ in period 0 to $0F_1'$ in period 1. In period 1, all firms can technically produce more output for each level of input, relative to what was possible in period 0. An example of technical change

is the installation of a new boiler for a coal-fired power plant that extends the plant productivity potential beyond previous limits.[6]

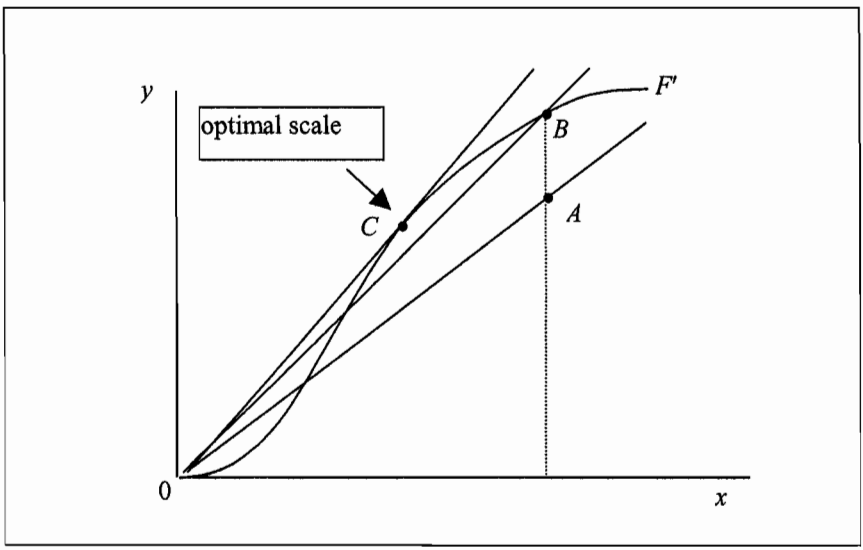

Figure 1.2 Productivity, Technical Efficiency and Scale Economies

When we observe that a firm has increased its productivity from one year to the next, the improvement need not have been from efficiency improvements alone, but may have been due to technical change or the exploitation of scale economies or from some combination of these three factors.

Up to this point, all discussion has involved physical quantities and technical relationships. We have not discussed issues such as costs or profits. If information on prices is available, and a behavioural assumption, such as cost minimisation or profit maximisation, is appropriate, then performance measures can be devised which incorporate this information. In such cases it is possible to consider **allocative efficiency,** in addition to technical efficiency. Allocative efficiency in input selection involves selecting that mix of inputs (e.g., labour and capital) that produces a given quantity of output at minimum cost (given the input prices which prevail). Allocative and technical efficiency combine to provide an overall economic efficiency measure.[7]

[6] This is an example of embodied technical change, where the technical change is embodied in the capital input. Disembodied technical change is also possible. One such example, is that of the introduction of legume/wheat crop rotations in agriculture in recent decades.

[7] In the case of a multiple-output industry, allocative efficiency in output mix may also be considered.

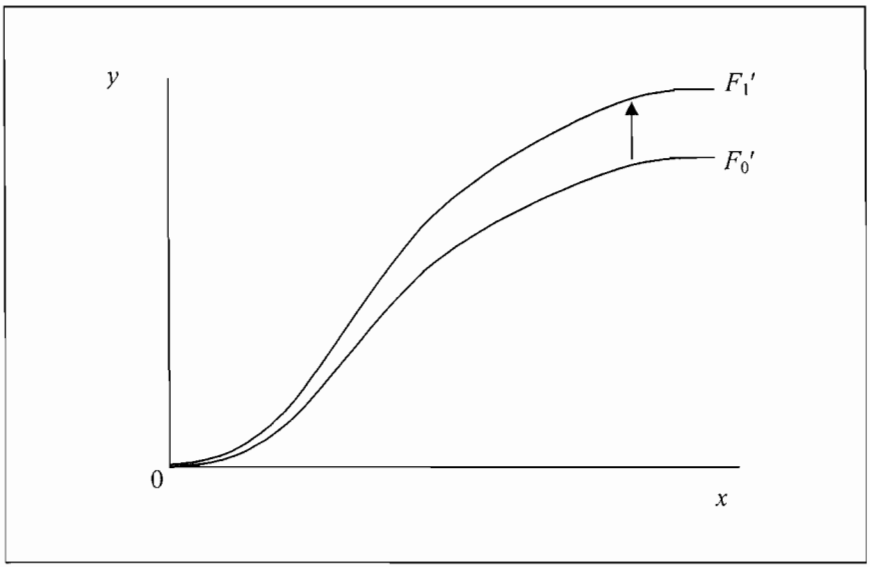

Figure 1.3 Technical Change Between Two Periods

Now that we are armed with this handful of informal definitions we briefly describe the layout of the book and the principal methods that we consider in subsequent chapters.

1.3 Overview of Methods

There are essentially four major methods discussed in this book:

1. least-squares econometric production models;
2. total factor productivity (TFP) indices;
3. data envelopment analysis (DEA); and
4. stochastic frontiers.

The first two methods are most often applied to aggregate time-series data and provide measures of technical change and/or TFP. Both of these methods assume all firms are technically efficient. Methods 3 and 4, on the other hand, are most often applied to data on a sample of firms (at one point in time) and provide measures of relative efficiency among those firms. Hence these latter two methods do not assume that all firms are technically efficient. However, multilateral TFP indices can also be used to compare the relative productivity of a group of firms at one point in time. Also DEA and stochastic frontiers can be used to measure both technical change and efficiency change, if panel data are available.

Thus we see that the above four methods can be grouped according to whether they recognise inefficiency or not. An alternative way of grouping the methods is to note that methods 1 and 4 involve the econometric estimation of parametric functions, while methods 2 and 3 do not. These two groups may therefore be termed "parametric" and "non-parametric" methods, respectively. These methods may also be distinguished in several other ways, such as by their data requirements, their behavioural assumptions and by whether or not they recognise random errors in the data (i.e. noise). These differences are discussed in later chapters.

1.4 Outline of Chapters

Summaries of the contents of the remaining 11 chapters are provided below.

Chapter 2. Review of Production Economics: This is a review of production economics at the level of an upper-undergraduate microeconomics course. It includes a discussion of the various ways in which one can provide a functional representation of a production technology, such as production, cost, revenue and profit functions, including information on their properties and dual relationships. We also review a variety of production economics concepts such as elasticities of substitution and returns to scale.

Chapter 3. Productivity and Efficiency Measurement Concepts: Here we describe how one can alternatively use set constructs to define production technologies analogous to those described using functions in Chapter 2. This is done because it provides a more natural way of dealing with multiple output production technologies, and allows us to introduce the concept of a distance function, which helps us define a number of our efficiency measurement concepts, such as technical efficiency. We also provide formal definitions of concepts such as technical efficiency, allocative efficiency, scale efficiency, technical change and total factor productivity (TFP) change.

Chapter 4. Index Numbers and Productivity Measurement: In this chapter we describe the familiar Laspeyres and Paasche index numbers, which are often used for price index calculations (such as a consumer price index). We also describe Tornqvist and Fisher indices and discuss why they may be preferred when calculating indices of input and output quantities and TFP. This involves a discussion of the economic theory that underlies various index number methods, plus a description of the various axioms that index numbers should ideally possess. We also cover a number of related issues such as chaining in time series comparisons and methods for dealing with transitivity violations in spatial comparisons.

Chapter 5. Data and Measurement Issues: In this chapter we discuss the very important topic of data set construction. We discuss a range of issues relating to the collection of data on inputs and outputs, covering topics such

as quality variations; capital measurement; cross-sectional and time-series data; constructing implicit quantity measures using price deflated value aggregates; aggregation issues, international comparisons; environmental differences; overheads allocation; plus many more. The index number concepts introduced in Chapter 4 are used regularly in this discussion.

Chapter 6. Data Envelopment Analysis: In this chapter we provide an introduction to DEA, the mathematical programming approach to the estimation of frontier functions and the calculation of efficiency measures. We discuss the basic DEA models (input- and output- orientated models under the assumptions of constant returns to scale and variable returns to scale) and illustrate these methods using simple numerical examples.

Chapter 7. Additional Topics on Data Envelopment Analysis: Here we extend our discussion of DEA models to include the issues of allocative efficiency; short run models; environmental variables; the treatment of slacks; super-efficiency measures; weights restrictions; and so on. The chapter concludes with a detailed empirical application.

Chapter 8. Econometric Estimation of Production Technologies: In this chapter we provide an overview of the main econometric methods that are used for estimating economic relationships, with an emphasis on production and cost functions. Topics covered include selection of functional form; alternative estimation methods (ordinary least squares, maximum likelihood, nonlinear least squares and Bayesian techniques); testing and imposing restrictions from economic theory; and estimating systems of equations. Even though the econometric models in this chapter implicitly assume no technical inefficiency, much of the discussion here is also useful background for the stochastic frontier methods discussed in the following two chapters. Data on rice farmers in the Philippines is used to illustrate a number of models.

Chapter 9. Stochastic Frontier Analysis: This is an alternative approach to the estimation of frontier functions using econometric techniques. It has advantages over DEA when data noise is a problem. The basic stochastic frontier model is introduced and illustrated using a simple example. Topics covered include maximum likelihood estimation, efficiency prediction and hypothesis testing. The rice farmer data from Chapter 8 is used to illustrate a number of models.

Chapter 10. Additional Topics on Stochastic Frontier Analysis: In this chapter we extend the discussion of stochastic frontiers to cover topics such as allocative efficiency, panel data models, the inclusion of environmental and management variables, risk modeling and Bayesian methods. The rice farmer data from Chapter 8 is used to illustrate a number of models.

Chapter 11. The Calculation and Decomposition of Productivity Change using Frontier Methods: In this chapter we discuss how one may use frontier methods (such as DEA and stochastic frontiers) in the analysis of panel data for the purpose of measuring TFP growth. We discuss how the TFP measures may be decomposed into technical efficiency change and technical change. The chapter concludes with a detailed empirical application using the rice farmer data from Chapter 8, which raises various topics including the effects of data noise, shadow prices and aggregation.

Chapter 12. Conclusions.

1.5 What is Your Economics Background?

When writing this book we had two groups of readers in mind. The first group contains postgraduate economics majors who have recently completed a graduate course on microeconomics, while the second group contains people with less knowledge of microeconomics. This second group might include undergraduate students, MBA students and researchers in industry and government who do not have a strong economics background (or who did their economics training a number of years ago). The first group may quickly review Chapters 2 and 3. The second group of readers should read Chapters 2 and 3 carefully. Depending on your background, you may also need to supplement your reading with some of the reference texts that are suggested in these chapters.

2. REVIEW OF PRODUCTION ECONOMICS

2.1 Introduction

This chapter reviews key economic concepts needed for a proper understanding of efficiency and productivity measurement. To make the chapter accessible we have chosen to use functions and graphs, rather than sets,[1] to describe the technological possibilities faced by firms. To further simplify matters, we assume i) the production activities of the firm take place in a single period, ii) the prices of all inputs and outputs are known with certainty, and iii) the firm is technically efficient in the sense that it uses its inputs to produce the maximum outputs that are technologically feasible (this last assumption is relaxed in Chapter 3). In all these respects, our review of production economics is similar to that found in most undergraduate economics textbooks.

We begin, in Section 2.2, by showing how the production possibilities of single-output firms can be represented using production functions. We explain some of the properties of these functions (eg., monotonicity) and define associated quantities of economic interest (eg., elasticities of substitution). In Section 2.3, we show how the production possibilities of *multiple*-output firms can be represented using transformation functions. However, this section is kept brief, not least because transformation functions can be viewed as special cases of the distance functions discussed in detail in Chapter 3. In Section 2.4, we show how multiple-output technologies can also be represented using cost functions. We discuss the properties of these functions and show how they can be used to quickly and easily

[1] Set representations of production technologies are discussed in Chapter 3.

derive input demand functions (using Shephard's Lemma). In Section 2.5, we briefly consider an alternative but less common representation of the production technology, the revenue function. Finally, in Section 2.6, we discuss the profit function. Among other things, we show that profit maximisation implies both cost minimisation and revenue maximisation.

Much of the material presented in this chapter is drawn from the microeconomics textbooks by Call and Holahan (1983), Chambers, (1988), Beattie and Taylor (1985), Varian (1992) and Henderson and Quandt (1980). More details are available in these textbooks, and almost any other microeconomics textbooks used in undergraduate economics classes.

2.2 Production Functions

Consider a firm that uses amounts of N inputs (eg., labour, machinery, raw materials) to produce a single output. The technological possibilities of such a firm can be summarised using the production function[2]

$$q = f(\mathbf{x}) \tag{2.1}$$

where q represents output and $\mathbf{x} = (x_1, x_2, ..., x_N)'$ is an $N \times 1$ vector of inputs. Throughout this chapter we assume these inputs are within the effective control of the decision maker. Other inputs that are outside the control of the decision maker (eg., rainfall) are also important, but, for the time being, it is convenient to subsume them into the general structure of the function $f(.)$. A more explicit treatment of these variables is provided in Section 10.6.

2.2.1 Properties

Associated with the production function 2.1 are several properties that underpin much of the economic analysis in the remainder of the book. Principal among these are (eg., Chambers, 1988):

F.1 *Nonnegativity:* The value of $f(\mathbf{x})$ is a finite, non-negative, real number.

F.2 *Weak Essentiality:* The production of positive output is impossible without the use of at least one input.

F.3 *Nondecreasing in* $\mathbf{x}$: (or *monotonicity*) Additional units of an input will not decrease output. More formally, if $\mathbf{x}^0 \geq \mathbf{x}^1$ then

[2] Most economics textbooks refer to the technical relationship between inputs and output as a production *function* rather than a production *frontier*. The two terms can be used interchangeably. The efficiency measurement literature tends to use the term *frontier* to emphasise the fact that the function gives the *maximum* output that is technologically feasible.

$f(\mathbf{x}^0) \geq f(\mathbf{x}^1)$. If the production function is continuously differentiable, monotonicity implies all marginal products are non-negative.

F.4 *Concave in* **x**: Any linear combination of the vectors $\mathbf{x}^0$ and $\mathbf{x}^1$ will produce an output that is no less than the same linear combination of $f(\mathbf{x}^0)$ and $f(\mathbf{x}^1)$. Formally[3], $f(\theta\mathbf{x}^0 + (1-\theta)\mathbf{x}^1) \geq \theta f(\mathbf{x}^0) + (1-\theta) f(\mathbf{x}^1)$ for all $0 \leq \theta \leq 1$. If the production function is continuously differentiable, concavity implies all marginal products are non-increasing (i.e., the well-known law of diminishing marginal productivity).

These properties are not exhaustive, nor are they universally maintained. For example, the monotonicity assumption is relaxed in cases where heavy input usage leads to *input congestion* (eg., when labour is hired to the point where "too many cooks spoil the broth"), and the weak essentiality assumption is usually replaced by a stronger assumption in situations where *every* input is essential for production.

To illustrate some of these ideas, Figure 2.1 depicts a production function defined over a single input, x. Notice that

- for the values of x represented on the horixontal axis, the values of q are all non-negative and finite real numbers. Thus, the function satisfies the non-negativity property F.1.
- the function passes through the origin, so it satisfies property F.2.
- the marginal product[4] of x is positive at all points between the origin and point G, implying the monotonicity property F.3 is satisfied at these points. However, monotonicity is violated at all points on the curved segment GR.
- as we move along the production function from the origin to point D, the marginal product of x increases. Thus, the concavity property F.4 is violated at these points. However, concavity is satisfied at all points on the curved segment DR.

In summary, the production function depicted in Figure 2.1 violates the concavity property in the region 0D and violates the monotonicity property in the region GR. However, it is consistent with all properties along the curved segment between points D and G – we refer to this as the *economically-feasible region* of production. Within this region, the point E is the point at which the average product[5] is maximised. We refer to this point as the *point of optimal scale* (of operations).

---

[3] For readers who are unfamiliar with vector algebra, when we pre-multiply a vector by a scalar we simply multiply every element of the vector by the scalar. For example, if $\mathbf{x} = (x_1, x_2, ..., x_N)'$ then $\theta\mathbf{x} = (\theta x_1, \theta x_2, ..., \theta x_N)'$.

[4] Graphically, the marginal product at a point is the slope of the production function at that point.

[5] In the case of a single-input production function the average product is AP = q/x. Graphically, the average product at a point is given by the slope of the ray that passes through the origin and that point.

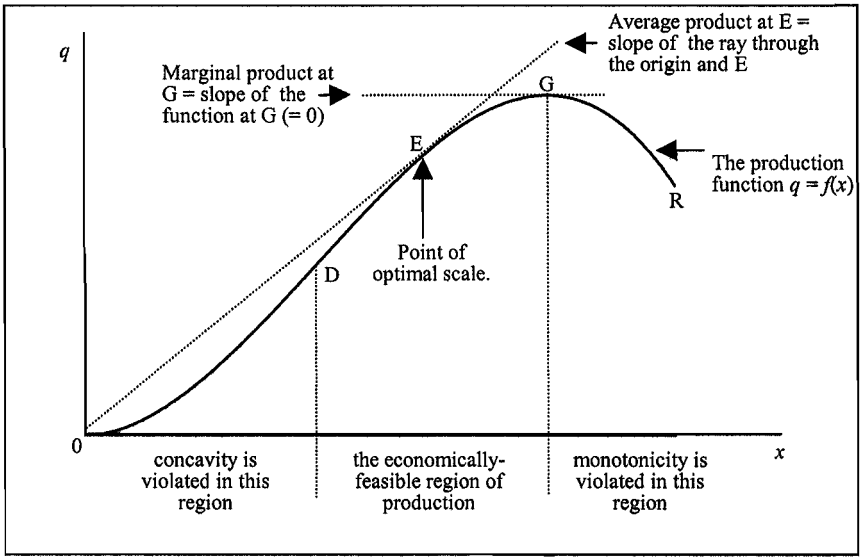

Figure 2.1 Single-Input Production Function

Extending this type of graphical analysis to the multiple-input case is difficult, not least because it is difficult to draw diagrams in more than two dimensions[6]. In such cases, it is common practice to plot the relationship between two of the variables while holding all others fixed. For example, in Figure 2.2 we consider a two-input production function and plot the relationship between the inputs x_1 and x_2 while holding output fixed at the value q^0. We also plot the relationship between the two inputs when output is fixed at the values q^1 and q^2, where $q^2 > q^1 > q^0$. The curves in this figure are known as output *isoquants*. If properties F.1 to F.4 are satisfied, these isoquants are non-intersecting functions that are convex to the origin, as depicted in Figure 2.2. The slope of the isoquant is known as the *marginal rate of technical substitution (MRTS)* – it measures the rate at which x_1 must be substituted for x_2 in order to keep output at its fixed level.

An alternative representation of a two-input production function is provided in Figure 2.3. In this figure, the lowest of the four functions, $q = f(x_1 | x_2 = x_2^0)$, plots the relationship between q and x_1 while holding x_2 fixed at the value x_2^0. The other functions plot the relationship between q and x_1 when x_2 is fixed at the values x_2^1, x_2^2, x_2^3 and x_2^4, where $x_2^4 > x_2^3 > x_2^2 > x_2^1 > x_2^0$.

[6] Some 3D representations of production functions can be found in Beattie and Taylor (1985).

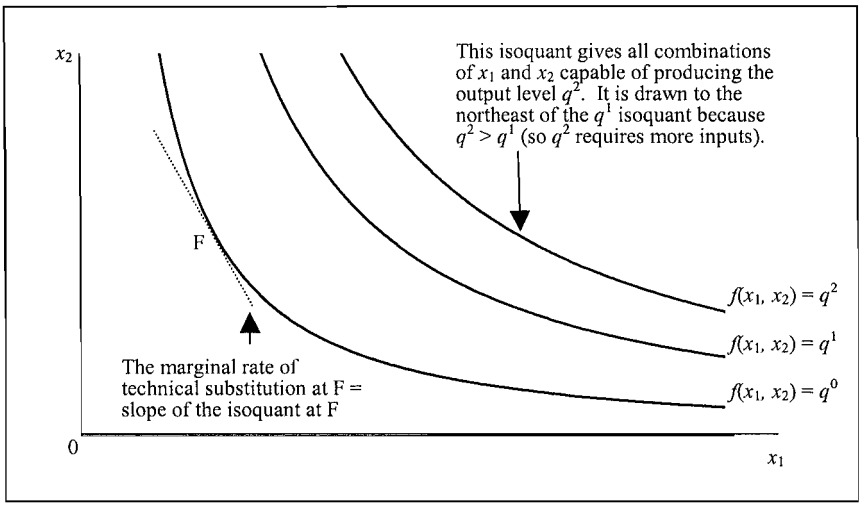

Figure 2.2 Output Isoquants

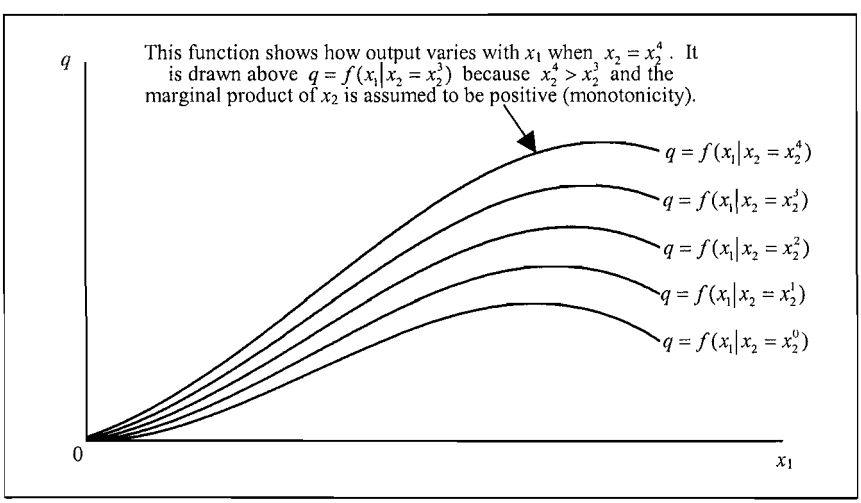

Figure 2.3 A Family of Production Functions

2.2.2 Quantities of Interest

If the production function 2.1 is twice-continuously differentiable we can use calculus to define a number of economic quantities of interest. For example, two quantities we have already encountered are the *marginal product,*

$$\mathrm{MP}_n = \frac{\partial f(\mathbf{x})}{\partial x_n}, \tag{2.2}$$

and the *marginal rate of technical substitution*:

$$\mathrm{MRTS}_{nm} = \frac{\partial x_n(x_1, \ldots, x_{n-1}, x_{n+1}, \ldots, x_N)}{\partial x_m} = -\frac{\mathrm{MP}_m}{\mathrm{MP}_n}. \tag{2.3}$$

In equation 2.3, $x_n(x_1, \ldots, x_{n-1}, x_n, \ldots, x_N)$ is an implicit function telling us how much of x_n is required to produce a fixed output when we use amounts $x_1, \ldots, x_{n-1}, x_n, \ldots, x_N$ of the other inputs[7]. Related concepts that do not depend on units of measurement are the *output elasticity*,

$$\mathrm{E}_n = \frac{\partial f(\mathbf{x})}{\partial x_n}\frac{x_n}{f(\mathbf{x})}, \tag{2.4}$$

and the *direct elasticity of substitution*:

$$\mathrm{DES}_{nm} = \frac{d(x_m / x_n)}{d(\mathrm{MP}_n / \mathrm{MP}_m)}\frac{\mathrm{MP}_n / \mathrm{MP}_m}{x_m / x_n}. \tag{2.5}$$

In the two-input case the DES is usually denoted σ.

Recall from Figure 2.2 that the MRTS measures the *slope* of an isoquant. The DES measures the percentage change in the input ratio relative to the percentage change in the MRTS, and is a measure of the *curvature* of the isoquant. To see this, consider movements along the isoquants depicted in Figure 2.4. In panel (a), an infinitesimal movement from one side of point A to the other results in an infinitesimal change in the input ratio but an infinitely large change in the MRTS, implying $\sigma \equiv \mathrm{DES}_{12} = 0$. Thus, in the case of a right-angled isoquant, an efficient firm must use its inputs in fixed proportions (i.e., no substitution is possible). In panel (c), a movement from D to E results in a large percentage change in the input ratio but leaves the MRTS unchanged, implying $\sigma = \infty$. In this case, the isoquant is a straight line and inputs are perfect substitutes. An intermediate (and more common) case is depicted in panel (b).

---

[7] For example, in the two-input case the implicit function $x_2(x_1)$ must satisfy $q^0 = f(x_1, x_2(x_1))$ where q^0 is a fixed value. Incidentally, differentiating both sides of this expression with respect to x_1 (and rearranging) we can show that MRTS is the negative of the ratio of the two marginal products (i.e., equation 2.3).

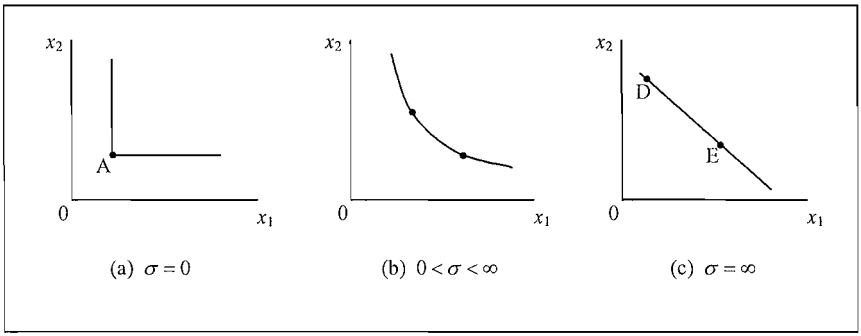

Figure 2.4 Elasticities of Substitution

In the multiple-input case it is possible to define at least two other elasticities of substitution – the *Allen partial elasticity of substitution* (AES) and the *Morishima elasticity of substitution* (MES). The DES is sometimes regarded as a short-run elasticity because it measures substitutability between x_n and x_m while holding all other inputs fixed (economists use the term "short-run" to refer to time horizons so short that at least one input is fixed). The AES and MES are long-run elasticities because they allow all inputs to vary. When there are only two inputs DES = AES. For more details see Chambers (1988, pp. 27-36).

The marginal product given by equation 2.2 measures the output response when one input is varied and all other inputs are held fixed. However, we are often interested in measuring output response when all inputs are varied simultaneously. If a proportionate increase in all inputs results in a *less than proportionate* increase in output (eg., doubling all inputs results in less than twice as much output) then we say the production function exhibits *decreasing returns to scale* (DRS). If a proportionate increase in inputs results in the *same* proportionate increase in output (eg., doubling all inputs results in exactly twice as much output) the production function is said to exhibit constant returns to scale (CRS). Finally, if a proportionate increase inputs leads to a *more than proportionate* increase in output the production function exhibits increasing returns to scale (IRS). Mathematically, if we scale all inputs by an amount $k > 1$ then

$$f(k\mathbf{x}) < kf(\mathbf{x}) \quad \Leftrightarrow \quad \text{DRS}, \tag{2.6}$$

$$f(k\mathbf{x}) = kf(\mathbf{x}) \quad \Leftrightarrow \quad \text{CRS}, \tag{2.7}$$

and

$$f(k\mathbf{x}) > kf(\mathbf{x}) \quad \Leftrightarrow \quad \text{IRS}. \tag{2.8}$$

There are many reasons why firms may experience different returns to scale. For example, a firm may exhibit IRS if the hiring of more staff permits some specialisation of labour, but may eventually exhibit DRS if it becomes so large that management is no longer able to exercise effective control over the production

process. Firms that can replicate *all* aspects of their operations exhibit CRS. Firms operating in regions of IRS are sometimes regarded as being too small, while firms operating in regions of DRS are sometimes regarded as being too large. In business and government, these considerations sometimes give rise to mergers, acquisitions, decentralisation, downsizing, and other changes in organisational structure.

In practice, a widely-used measure[8] of returns to scale is the *elasticity of scale* (or *total elasticity of production*),

$$\varepsilon = \frac{df(k\mathbf{x})}{dk}\frac{k}{f(k\mathbf{x})}\bigg|_{k=1} = \sum_{n=1}^{N} \mathrm{E}_n \tag{2.9}$$

where E_n is the output elasticity given by equation 2.4. The production function exhibits locally DRS, CRS or IRS as the elasticity of scale is less than, equal to, or greater than 1. We use the term "locally" because, like all measures derived using differential calculus, this particular measure only tells us what happens to output when inputs are scaled up or down by an infinitesimally small amount.

2.2.3 An Example

To illustrate the computation of marginal products and elasticities, consider the two-input Cobb-Douglas[9] production function

$$q = 2x_1^{0.5}x_2^{0.4}. \tag{2.10}$$

This simple production function will be used for all but one of the numerical examples in this chapter. For this production technology,

$$\mathrm{MP}_1 = \frac{\partial q}{\partial x_1} = x_1^{-0.5}x_2^{0.4} \tag{2.11}$$

$$\mathrm{MP}_2 = \frac{\partial q}{\partial x_2} = 0.8x_1^{0.5}x_2^{-0.6} \tag{2.12}$$

$$\mathrm{E}_1 = \frac{\partial q}{\partial x_1}\frac{x_1}{q} = \left(x_1^{-0.5}x_2^{0.4}\right)\left(\frac{x_1}{2x_1^{0.5}x_2^{0.4}}\right) = 0.5 \tag{2.13}$$

and $$\mathrm{E}_2 = \frac{\partial q}{\partial x_2}\frac{x_2}{q} = \left(0.8x_1^{0.5}x_2^{-0.6}\right)\left(\frac{x_2}{2x_1^{0.5}x_2^{0.4}}\right) = 0.4 \tag{2.14}$$

---

[8] Another convenient measure of returns to scale is the degree of homogeneity of the production function. A function is said to be *homogenous of degree r* if $f(k\mathbf{x}) = k^r f(\mathbf{x})$ for all $k > 0$. Thus, a production function will exhibit local DRS, CRS or IRS as the degree of homogeneity is less than, equal to, or greater than 1. A function that is homogeneous of degree 1 is said to be *linearly homogeneous*.

[9] The Cobb-Douglas form is just one of many functional forms used by economists to specify relationships between economic variables. Several functional forms are listed in Section 8.2.

Thus, the output elasticities do not vary with variations in input levels. This is a well-known and arguably restrictive property of *all* Cobb-Douglas production functions[10]. One important consequence is that the elasticity of scale is also constant:

$$\varepsilon = \sum_{n=1}^{2} \mathrm{E}_n = 0.5 + 0.4 = 0.9. \tag{2.15}$$

This elasticity is less than 1, implying the technology everywhere exhibits local DRS[11]. Finally, to calculate $\sigma \equiv \mathrm{DES}_{12}$, we note from equations 2.11 and 2.12 that

$$\frac{\mathrm{MP}_1}{\mathrm{MP}_2} = \frac{x_1^{-0.5} x_2^{0.4}}{0.8 x_1^{0.5} x_2^{-0.6}} = \frac{1}{0.8}\left(\frac{x_2}{x_1}\right), \tag{2.16}$$

or, after some algebra,

$$\frac{x_2}{x_1} = 0.8\left(\frac{\mathrm{MP}_1}{\mathrm{MP}_2}\right). \tag{2.17}$$

Finally,

$$\sigma = \frac{d(x_2/x_1)}{d(\mathrm{MP}_1/\mathrm{MP}_2)} \times \frac{\mathrm{MP}_1/\mathrm{MP}_2}{x_2/x_1} = 0.8 \times \frac{1}{0.8}\left(\frac{x_2}{x_1}\right) \times \left(\frac{x_1}{x_2}\right) = 1. \tag{2.18}$$

Thus, the direct elasticity of substitution is equal to 1. This is another restrictive property of *all* Cobb-Douglas production functions.

2.2.4 Short-Run Production Functions

We have already mentioned that economists use the term 'short-run' to refer to time horizons so short that some inputs must be treated as fixed (usually buildings and other forms of capital infrastructure). Conversely, the term 'long-run' is used to refer to time horizons long enough that all inputs can be regarded as variable. Until now we have been treating all inputs as variable. Thus, production functions such as 2.10 can be viewed as *long-run production functions*.

Short-run variants of long-run production functions are obtained by simply holding one or more inputs fixed. For example, consider the production function 2.10 and suppose the second input is fixed at the value $x_2 = 100$, at least in the short run. The resulting *short-run production function* is:

---

[10] More generally, for the Cobb-Douglas production function defined over N inputs, $q = Ax_1^{\beta_1} x_2^{\beta_2} \ldots x_N^{\beta_N}$, the output elasticities are $\mathrm{E}_n = \beta_n$.

[11] We can verify this by replacing x_1 and x_2 in equation 2.1 with kx_1 and kx_2 and observing what happens to output.

$$q = 2x_1^{0.5}100^{0.4} = 12.619x_1^{0.5}. \tag{2.19}$$

This function is depicted in Figure 2.5. Of course, at some time in the future the firm might find that the second input is temporarily fixed at another value, say x_2 = 150. In this case the short-run production function is

$$q = 14.841x_1^{0.5}. \tag{2.20}$$

This function is also depicted in Figure 2.5. If we repeat this exercise a number of times we will eventually construct a family of short-run production functions, each of which can be seen to satisfy properties F.1 to F.4. As a group, this family could be viewed as a long-run production function (because it depicts the production possibilities of the firm as both inputs vary).[12]

2.3 Transformation Functions

We can generalise the production function concept to the case of a firm that produces more than one output. Specifically, the technological possibilities of a firm that uses N inputs to produce M outputs can be summarized by the transformation function:

$$T(\mathbf{x}, \mathbf{q}) = 0, \tag{2.21}$$

where $\mathbf{q} = (q_1, q_2, ..., q_M)'$ is an $M \times 1$ vector of outputs. A special case of a transformation function is the production function 2.1 expressed in implicit form:

$$T(\mathbf{x}, q) = q - f(\mathbf{x}) = 0. \tag{2.22}$$

Thus, it should be no surprise that transformation functions have properties that are analogous to properties F.1 to F.5. In addition, if they are twice-continuously differentiable we can use calculus to derive expressions for economic quantities of interest, as we did in Section 2.2.2. Details are not provided in this chapter, for two reasons. First, we can view transformation functions as special cases of the distance functions discussed in detail in Chapter 3. Second, most applied economists analyse multiple-output technologies in ways that do not involve the specification of transformation functions or their properties. Some simply aggregate the outputs into a single measure using the index number methods discussed in Chapter 4 (and then use the production function to summarise technically-feasible production plans). Others make use of price information and represent the technology using the cost, revenue and profit functions discussed below.

[12] The production functions depicted in Figure 2.3 could also be viewed as a family of short-run production functions.

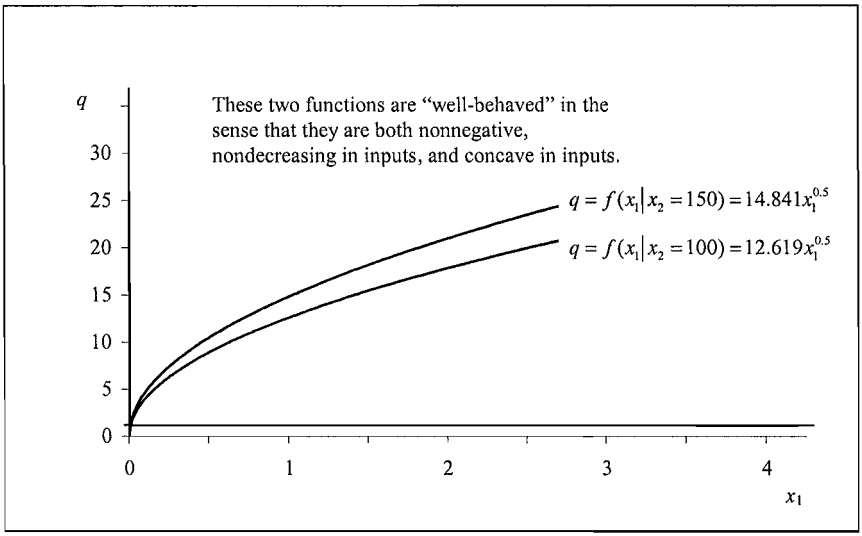

Figure 2.5 Short-Run Production Functions

2.4 Cost Functions

Until now we have been concerned with the physical relationships between inputs and outputs. In this section we look at how firms decide on the mix of inputs they wish to use. The most common assumption is that firms make these decisions in order to minimise costs.

Consider the case of a multiple-input multiple-output firm that is so small relative to the size of the market that it has no influence on input prices – it must take these prices as given. Such a firm is said to be *perfectly competitive* in input markets. Mathematically, the cost minimisation problem for this firm can be written

$$c(\mathbf{w},\mathbf{q}) = \min_{\mathbf{x}} \ \mathbf{w}'\mathbf{x} \quad \text{such that} \quad T(\mathbf{q},\mathbf{x}) = 0. \tag{2.23}$$

where $\mathbf{w} = (w_1, w_2, ..., w_N)'$ is a vector of input prices. The right-hand side of this equation says "search over all technically feasible input-output combinations and find the input quantities that minimise the cost of producing the output vector $\mathbf{q}$".[13] We have used the notation $c(\mathbf{w},\mathbf{q})$ on the left-hand side to emphasise that this minimum cost value varies with variations in $\mathbf{w}$ and $\mathbf{q}$.

[13] For readers who are unfamiliar with vector algebra, we should explain that the term $\mathbf{w}'\mathbf{x}$ is the *inner product* of the vectors $\mathbf{w}$ and $\mathbf{x}$. It is a compact way of writing the sum of the products of the corresponding elements. That is, $\mathbf{w}'\mathbf{x} = w_1x_1 + w_2x_2 + ... + w_Nx_N$ = cost.

2.4.1 An Example

As an example of a cost minimisation problem, consider a single-output two-input firm having the production function $q = 2x_1^{0.5}x_2^{0.4}$ (i.e., the production function used in Section 2.2.3). In this case the cost minimisation problem can be written[14]

$$c(w_1, w_2, q) = \min_{x_1, x_2} \; w_1x_1 + w_2x_2 \quad \text{such that} \quad x_2 - 0.177x_1^{-1.25}q^{2.5} = 0 \tag{2.24}$$

or, substituting for x_2,

$$c(w_1, w_2, q) = \min_{x_1} \; w_1x_1 + 0.177w_2x_1^{-1.25}q^{2.5}. \tag{2.25}$$

Minimising the function $w_1x_1 + 0.177w_2x_1^{-1.25}q^{2.5}$ with respect to x_1 is a simple exercise in differential calculus. We simply take the first derivative with respect to x_1 and set it to zero:[15]

$$w_1 - 0.221w_2x_1^{-2.25}q^{2.5} = 0. \tag{2.26}$$

Solving for x_1 we obtain the *conditional*[16] *input demand function*:

$$x_1(w_1, w_2, q) = 0.511w_1^{-0.444}w_2^{0.444}q^{1.111}. \tag{2.27}$$

Substituting equation 2.27 back into the technology constraint yields a second conditional input demand function:

$$x_2(w_1, w_2, q) = 0.409w_1^{0.556}w_2^{-0.556}q^{1.111}. \tag{2.28}$$

Finally, the cost function is:

$$c(w_1, w_2, q) = w_1x_1(w_1, w_2, q) + w_2x_2(w_1, w_2, q) = 0.92w_1^{0.556}w_2^{0.444}q^{1.111}. \tag{2.29}$$

An interesting property of this cost function is that it has the same functional form as the production function 2.10 (i.e., Cobb-Douglas). This property is shared by *all* Cobb-Douglas production and cost functions. Such functions are said to be *self-dual.*

We can gain some insights into the properties of the cost function 2.29 by computing the cost-minimising input demands (and associated minimum costs) at different values of the right-hand-side variables. For example, when we substitute

---

[14] The technology constraint $x_2 - 0.177x_1^{-1.25}q^{2.5} = 0$ is obtained by simply rearranging $q = 2x_1^{0.5}x_2^{0.4}$. .

[15] It is straightforward to show that the second order condition for a maximum (i.e. second-order derivative less than zero) is also satisfied for all non-negative values of **w** and q. Similar second-order conditions are satisfied for other optimisation problems considered in this chapter.

[16] This terminology derives from the fact that the input demands are *conditional* on the value of output.

the values $(w_1, w_2, q) = (150, 1, 10)$ into equations 2.27 to 2.29 we find $(x_1, x_2, c) = (0.71, 85.63, 192.62)$. That is, a (minimum) cost of 192.62 is incurred when the firm uses amounts $x_1 = 0.71$ and $x_2 = 85.63$. If we double the two input prices we find that the minimum cost doubles to $c = 385.23$ while input demands remain unchanged. Finally, if we keep input prices at $(w_1, w_2) = (150, 1)$ but increase output to $q = 15$ we find $(x_1, x_2, c) = (1.12, 134.36, 302.23)$. These computations confirm that our cost function exhibits some familiar and commonsense properties – it is nondecreasing and linearly homogeneous in prices, and nondecreasing in output. Other properties are listed in Section 2.4.2 below.

Several aspects of this numerical example are depicted graphically in Figure 2.6. To construct this figure we have rewritten the cost function $c = w_1x_1 + w_2x_2$ in the form $x_2 = (c/w_2) - (w_1/w_2)x_2$. This is the equation of an *isocost line* – a straight line with intercept c/w_2 and slope $-(w_1/w_2)$ that gives all input combinations that cost c. In Figure 2.6 we plot two isocost lines, both with slope $-(w_1/w_2) = -150$. We also plot two isoquants corresponding to $q = 10$ and $q = 15$. Note that the isocost line closest to the origin is tangent to the $q = 10$ isoquant at $(x_1, x_2) = (0.71, 85.63)$. The second isocost line is tangent to the $q = 15$ isoquant at $(x_1, x_2) = (1.12, 134.36)$. These are the two cost-minimising solutions computed above.

2.4.2 Properties

Irrespective of the properties of the production technology, the cost function satisfies the following properties:

C.1	*Nonnegativity:*	Costs can never be negative.
C.2	*Nondecreasing in* **w**:	An increase in input prices will not decrease costs. More formally, if $\mathbf{w}^0 \geq \mathbf{w}^1$ then $c(\mathbf{w}^0, \mathbf{q}) \geq c(\mathbf{w}^1, \mathbf{q})$.
C.3	*Nondecreasing in* **q**:	It costs more to produce more output. That is, if $\mathbf{q}^0 \geq \mathbf{q}^1$ then $c(\mathbf{w}, \mathbf{q}^0) \geq c(\mathbf{w}, \mathbf{q}^1)$.
C.4	*Homogeneity:*	Multiplying all input prices by an amount $k > 0$ will cause a k–fold increase in costs (eg., doubling all input prices will double cost). Mathematically, $c(k\mathbf{w}, \mathbf{q}) = kc(\mathbf{w}, \mathbf{q})$ for $k > 0$.
C.5	*Concave in* **w**:	$c(\theta\mathbf{w}^0 + (1-\theta)\mathbf{w}^1, \mathbf{q}) \geq \theta c(\mathbf{w}^0, \mathbf{q}) + (1-\theta)c(\mathbf{w}^1, \mathbf{q})$ for all $0 \leq \theta \leq 1$. This statement is not very intuitive. However, an important implication of the property is that input demand functions cannot slope upwards.

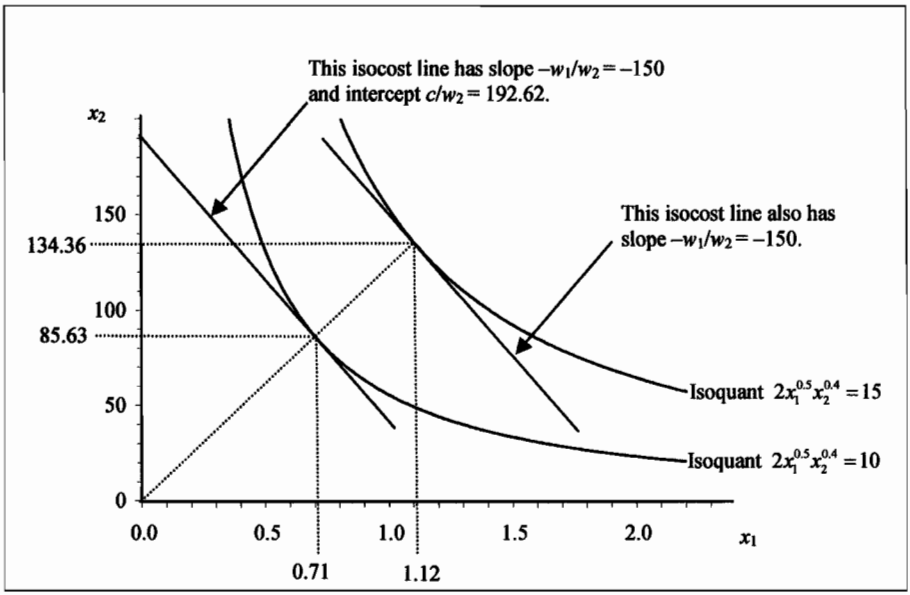

Figure 2.6 Cost Minimisation

These properties of the cost function are used by economists in at least three important ways. First, in the absence of changes in technology or market structure, evidence that one or more properties are violated can be regarded as evidence that a firm is not minimising costs. For example, we might use accounting data to check that a proportionate increase in all nominal input prices (eg., through currency movements or inflation) has resulted in the same proportionate increase in nominal costs. This will indicate whether the homogeneity property C.4 holds. Second, they can be used to establish qualitative results concerning changes in market structure or government policy. For example, the concavity property can be used to show that average costs under an input-price stabilisation scheme are no less than average costs under fluctuating prices. Finally, they can be used to obtain better econometric estimates of cost and conditional input demand functions. Econometric methods for incorporating some types of regularity properties into the estimation process are discussed in Chapter 8.

2.4.3 Deriving Conditional Input Demand Equations

In Section 2.4.1 we derived the conditional input demand equations 2.27 and 2.28 by explicitly solving the cost minimisation problem 2.24. Both equations were then used to construct the cost function 2.29. The simplicity of this example was largely due to our use of a two-input single-output Cobb-Douglas production function.

Unfortunately, the algebra quickly becomes unmanageable when we have more than a few inputs and outputs and/or we use a functional form that is less tractable than the Cobb-Douglas.

When dealing with multiple-input multiple-output technologies, it is usually more convenient (and common) to derive conditional input demand equations by working back from a well-behaved cost function. Specifically, if the cost function is twice-continuously differentiable then *Shephard's Lemma* says that:

$$x_n(\mathbf{w},\mathbf{q}) = \frac{\partial c(\mathbf{w},\mathbf{q})}{\partial w_n}. \tag{2.30}$$

This result has an important practical implication – once a well-behaved cost function has been specified or estimated econometrically, we can use Shephard's Lemma to quickly and easily obtain the conditional input demand equations. To illustrate, consider the cost function 2.29 derived in Section 2.4.1:

$$c(w_1,w_2,q) = w_1x_1(w_1,w_2,q) + w_2x_2(w_1,w_2,q) = 0.92w_1^{0.556}w_2^{0.444}q^{1.111}. \tag{2.29}$$

The first-order derivatives with respect to prices are

$$x_1(w_1,w_2,q) = 0.511w_1^{-0.444}w_2^{0.444}q^{1.111}. \tag{2.31}$$

and $$x_2(w_1,w_2,q) = 0.409w_1^{0.556}w_2^{-0.556}q^{1.111}. \tag{2.32}$$

These equations are identical to the input demand equations 2.27 and 2.28.

This approach, where Shephard's Lemma is used to derive input demand equations, is known as the *dual approach*. The approach used earlier, involving constrained minimisation of the cost function, is known as the *primal approach*. In practice, the dual approach is used much more widely than the primal approach, partly because it is easier, but also because (estimated) cost functions are often closer to hand than production functions.[17]

Finally, if the cost function is twice-continuously differentiable and satisfies properties C.1 to C.5, Shephard's Lemma can be used to show that conditional input demand functions have the properties:

D.1 *Nonnegativity*: $x_n(\mathbf{w}, \mathbf{q}) \geq 0$.

D.2 *Nonincreasing in* **w**: $\partial x_n(\mathbf{w},\mathbf{q}) / \partial w_n \leq 0$.

[17] Econometricians often find it easier to estimate cost functions than production functions, partly because price data is usually easier to obtain than quantity data, and partly because there are usually fewer econometric difficulties to deal with (e.g.,endogeneity is not usually an issue in cost function estimation because prices are usually exogenous).

D.3 *Nondecreasing in* $\mathbf{q}$: $\partial x_n(\mathbf{w},\mathbf{q})/\partial q_m \geq 0$.

D.4 *Homogeneity*: $x_n(k\mathbf{w}, \mathbf{q}) = x_n(\mathbf{w}, \mathbf{q})$ for $k > 0$.

D.5 *Symmetry*: $\partial x_n(\mathbf{w},\mathbf{q})/\partial w_m = \partial x_m(\mathbf{w},\mathbf{q})/\partial w_n$.

Some of these properties are evident in Figure 2.6. In particular, we can see that a proportionate change in input prices leaves the slopes of the isocost lines, and therefore the cost-minimising input quantities, unchanged (i.e., the input demands are homogeneous of degree zero in prices). We can also see that moving to a higher isoquant is associated with an increase in input usage (i.e., the input demand functions are nondecreasing in output).

2.4.4 The Short-Run Cost Function

Until now we have assumed that all inputs are variable, as they would be in the long run[18]. For this reason, the cost function $c(\mathbf{w},q)$ is sometimes known as a *variable* or *long-run cost function*. A useful variant of this function is obtained by assuming that a subset of inputs are fixed, as some inputs would be in the short run (eg., buildings). The resulting cost function is known as a *restricted* or *short-run cost function*.

Let the input vector $\mathbf{x}$ be partitioned as $\mathbf{x} = (\mathbf{x}_f, \mathbf{x}_v)$ where $\mathbf{x}_f$ and $\mathbf{x}_v$ are subvectors containing fixed and variable inputs respectively, and let the input price vector $\mathbf{w}$ be similarly partitioned as $\mathbf{w} = (\mathbf{w}_f, \mathbf{w}_v)$. Then the short-run cost minimisation problem can be written

$$c(\mathbf{w},\mathbf{q},\mathbf{x}_f) = \min_{\mathbf{x}_v} \ \mathbf{w}_v'\mathbf{x}_v + \mathbf{w}_f'\mathbf{x}_f \quad \text{such that} \quad T(\mathbf{q},\mathbf{x}) = 0. \tag{2.33}$$

Note that this problem only involves searching over values of the *variable* inputs. In every other respect, it is identical to the long-run cost minimisation problem 2.23. Thus, it is not surprising that $c(\mathbf{w},\mathbf{q},\mathbf{x}_f)$ satisfies properties C.1 to C.5 (although the nonnegativity property can be strengthened – the short-run function is *strictly positive* owing to the existence of fixed input costs). In addition, $c(\mathbf{w},\mathbf{q},\mathbf{x}_f) \geq c(\mathbf{w},\mathbf{q})$ (i.e., short-run costs are no less than long-run costs), and if $\mathbf{x}_f^0 \geq \mathbf{x}_f^1$ then $c(\mathbf{w},\mathbf{q},\mathbf{x}_f^0) \geq c(\mathbf{w},\mathbf{q},\mathbf{x}_f^1)$ (i.e., the function is nondecreasing in fixed inputs).

To illustrate these last two properties, suppose the second input in our Cobb-Douglas production function $q = 2x_1^{0.5}x_2^{0.4}$ is fixed. The short-run cost minimisation problem is

[18] The concepts of 'long run' and 'short run' are briefly discussed in Section 2.2.4.

$$c(w_1, w_2, q, x_2) = \min_{x_1} \; w_1 x_1 + w_2 x_2 \quad \text{such that} \quad q - 2x_1^{0.5} x_2^{0.4} = 0. \tag{2.34}$$

which is identical to the problem 2.24 except we are now only minimising over x_1 (and the technology constraint has been written in a slightly different way). The technology constraint can be solved for the short-run conditional input demand function:[19]

$$x_1(w_1, w_2, q, x_2) = 0.25 x_2^{-0.8} q^2 \tag{2.35}$$

so the short-run cost function is

$$c(w_1, w_2, q, x_2) = w_1 x_1(w_1, w_2, q, x_2) + w_2 x_2 = 0.25 w_1 x_2^{-0.8} q^2 + w_2 x_2. \tag{2.36}$$

Evaluating equations 2.35 and 2.36 at the point (w_1, w_2, q , x_2) = (150, 1, 10, 100) yields x_1 = 0.63 and c = 194.20. This minimum cost value is slightly higher than the minimum cost value of c = 192.62 we computed in Section 2.4.1 using the long-run cost function 2.29 (and the same values of w_1, w_2 and q). Repeating the exercise using (w_1, w_2, q , x_2) = (150, 1, 10 200) (i.e., when we double the amount of the fixed input but hold all other variables fixed) we find x_1 = 0.361 and c = 254.10 (i.e., minimum cost has increased).

2.4.5 Marginal and Average Costs

Associated with long-run and short-run cost functions are several concepts that are frequently used when discussing firm behaviour. For example, in the case of a single-output firm we can define

Short-run variable cost: $$\text{SVC} = \mathbf{w}_v' \mathbf{x}_v(\mathbf{w}, q, \mathbf{x}_f) \tag{2.37}$$

Short-run fixed cost: $$\text{SFC} = \mathbf{w}_f' \mathbf{x}_f \tag{2.38}$$

Short-run total cost: $$\text{STC} = \mathbf{w}_v' \mathbf{x}_v(\mathbf{w}, q, \mathbf{x}_f) + \mathbf{w}_f' \mathbf{x}_f \tag{2.39}$$

Short-run average variable cost: $$\text{SAVC} = \frac{\mathbf{w}_v' \mathbf{x}_v(\mathbf{w}, q, \mathbf{x}_f)}{q} \tag{2.40}$$

Short-run average cost: $$\text{SAC} = \frac{c(\mathbf{w}, q, \mathbf{x}_f)}{q} \tag{2.41}$$

[19] In this simple two-input example there are no input substitution possibilities (because x_2 is fixed). Thus, the conditional demand for x_1 does not depend on input prices. When there is more than one variable input, the short-run conditional input demand functions will usually be functions of variable input prices. For this reason, we use the more general notation $x_1(w_1, w_2, q, x_2)$ on the left-hand side of equation 2.35.

Short-run average fixed cost: $$\text{SAFC} = \frac{\mathbf{w}_f' \mathbf{x}_f}{q} \tag{2.42}$$

Short-run marginal cost: $$\text{SMC} = \frac{\partial c(\mathbf{w}, q, \mathbf{x}_f)}{\partial q} \tag{2.43}$$

Long-run total cost: $$\text{LTC} = c(\mathbf{w}, q) \tag{2.44}$$

Long-run average cost: $$\text{LAC} = \frac{c(\mathbf{w}, q)}{q} \tag{2.45}$$

Long-run marginal cost: $$\text{LMC} = \frac{\partial c(\mathbf{w}, q)}{\partial q} \tag{2.46}$$

All of these concepts should be self-evident. Note that we have not defined "long-run average variable cost" or "long-run average fixed cost" because all costs are variable in the long run.

To illustrate the nature of some of these quantities, suppose the production function is $q = 2x_1^{0.5}x_2^{0.4}$ and x_2 is fixed in the short run. Then:

Short-run variable cost $$\text{SVC} = 0.25 w_1 x_2^{-0.8} q^2 \tag{2.47}$$

Short-run fixed cost $$\text{SFC} = w_2 x_2 \tag{2.48}$$

Short-run total cost $$\text{STC} = 0.25 w_1 x_2^{-0.8} q^2 + w_2 x_2 \tag{2.49}$$

Short-run average cost: $$\text{SAC} = 0.25 w_1 x_2^{-0.8} q + w_2 x_2 q^{-1} \tag{2.50}$$

Short-run average variable cost: $$\text{SAVC} = 0.25 w_1 x_2^{-0.8} q \tag{2.51}$$

Short-run average fixed cost: $$\text{SAFC} = w_2 x_2 q^{-1} \tag{2.52}$$

Short-run marginal cost: $$\text{SMC} = 0.5 w_1 x_2^{-0.8} q \tag{2.53}$$

Long-run total cost: $$\text{LTC} = 0.92 w_1^{0.556} w_2^{0.444} q^{1.111} \tag{2.54}$$

Long-run average cost: $$\text{LAC} = 0.92 w_1^{0.556} w_2^{0.444} q^{0.111} \tag{2.55}$$

Long-run marginal cost: $$\text{LMC} = 1.022 w_1^{0.556} w_2^{0.444} q^{0.111} \tag{2.56}$$

These equations express various types of costs as functions of up to four variables. We can represent them graphically by holding all but one of the right-hand side variables fixed. For example, in Figure 2.7 we set $(w_1, w_2, x_2) = (150, 1\ 100)$ and plot SVC, SFC and STC against output. Notice that the STC function is simply the sum of the SVC and SFC functions, and that it is convex in q.

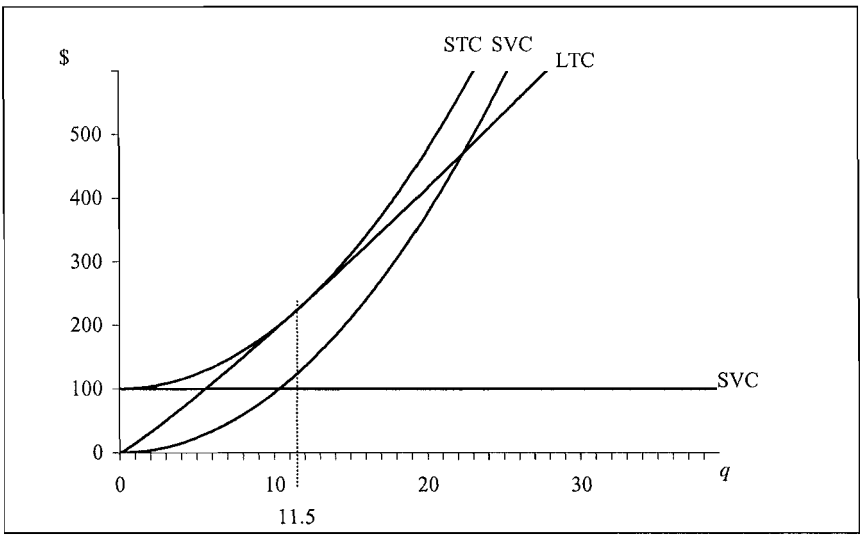

Figure 2.7 Long-Run and Short-Run Fixed, Variable and Total Costs

Some readers will notice that the STC function in Figure 2.7 is not typical of the STC functions depicted in most economics textbooks. For example, Beattie and Taylor (1985, p.176) draw an STC function that is concave in q when q is small, and convex in q when q is large. The difference between Figure 2.7 and the figure in Beattie and Taylor (1985) can be traced back to the properties of the underlying short-run production functions. Specifically, the short-run production function underpinning Figure 2.7 is the lower of the two "well-behaved" functions depicted in Figure 2.5, while the production function underpinning the textbook STC function has the shape depicted in Figure 2.1. The concave segment in the textbook STC function is associated with the region 0D in Figure 2.1, where the variable input exhibits *increasing* marginal productivity.

Finally, Figure 2.7 also plots the LTC curve. This LTC function is ever-so-slightly convex in q, a property that must be satisfied if there is to be a unique solution to the long-run profit maximisation problem (we revisit this issue in Section 2.6 below). It is also tangent to the STC function when $q = 11.5$. Recall that the STC function was obtained by setting $(w_1, w_2, x_2) = (150, 1\ 100)$. The conditional input demand equations 2.31 and 2.32 can be used to verify that the values $x_1 = 0.833$ and $x_2 = 100$ minimise the long-run cost of producing output $q = 11.5$ when prices are $w_1 = 150$ and $w_2 = 1$.

2.4.6 Economies of Scale and Scope

In Section 2.2.2 we used a single-output production function to define several economic quantities of interest, including a measure of returns to scale. Measures of returns to scale are also available in the multiple-output case, and they can be defined in terms of the cost function. For example, a measure of overall scale economies is

$$\varepsilon_c = \left[\sum_{m=1}^{M} \frac{\partial \ln c(\mathbf{w},\mathbf{q})}{\partial \ln q_m} \right]^{-1} \tag{2.57}$$

The firm will exhibit increasing, constant or decreasing returns to scale as ε_c is greater than, equal to, or less than one.

In the multiple-output case, it is also meaningful to consider the cost savings resulting from producing different numbers of outputs. Three measures of so-called *economies of scope* are:

$$S = \left[\sum_{m=1}^{M} c(\mathbf{w},q_m) \Big/ c(\mathbf{w},\mathbf{q}) \right] - 1 \tag{2.58}$$

$$S_m = \frac{c(\mathbf{w},q_m) + c(\mathbf{w},\mathbf{q}_{M-m}) - c(\mathbf{w},\mathbf{q})}{c(\mathbf{w},\mathbf{q})} \tag{2.59}$$

and $$S_{mn} = \frac{\partial^2 c(\mathbf{w},\mathbf{q})}{\partial q_m \partial q_n} \tag{2.60}$$

where $c(\mathbf{w},q_m)$ denotes the cost of producing the m-th output only; and $c(\mathbf{w},\mathbf{q}_{M-m})$ denotes the cost of producing all outputs except the m-th output. The measure defined by equation 2.58 is a measure of *global* economies of scope, and gives the proportionate change in costs if all outputs are produced separately – if $S > 0$ then it is best to produce all outputs as a group; if $S < 0$ then it is best to produce all outputs separately. The measure defined by equation 2.59 is a measure of *product-specific* economies of scope, and gives the proportionate change in costs if the m-th output is produced separately and all other outputs are produced as a group – if $S_m > 0$ then it is best to produce all outputs as a group; if $S_m < 0$ then it is best to produce the m-th output separately. Finally, the measure defined by equation 2.60 is another measure of product-specific economies of scope. It gives the change in the marginal cost of producing the m-th output with respect to a change in the production of the n-th output. The firm experiences economies of scope with respect to the n-th output if this derivative is negative. For more details and some empirical examples see Glass, McKilllop and Hyndman (1995) and Deller, Chicoine and Walzer (1988).

2.5 Revenue Functions

We have just seen how to determine the minimum cost of producing a given output vector **q**. A similar problem is that of determining the maximum revenue that can be obtained from a given input vector **x**. The function that gives us this maximum revenue is known as a *revenue function*. In this section we do little more than present the revenue function and its properties. We keep it brief for two reasons. First, the problem of maximising revenue exactly mirrors the problem of minimising cost, and we want to avoid repetition (in fact, in Section 2.6 we will see that the revenue and cost functions are both restricted variants of a profit function). Second, applied production economists use the revenue function much less frequently than the cost function (the revenue function is more widely used in macroeconomics and international trade where, for example, economists are interested in studying the maximum income a country can generate from a given resource endowment).

The revenue maximisation problem for a multiple-input multiple-output firm can be written

$$r(\mathbf{p},\mathbf{x}) = \max_{\mathbf{q}} \ \mathbf{p}'\mathbf{q} \quad \text{such that} \quad T(\mathbf{q},\mathbf{x}) = 0 \tag{2.61}$$

where $\mathbf{p} = (p_1, p_2, ..., p_M)'$ is a vector of output prices over which the firm has no influence (i.e., it is perfectly competitive in output markets). The revenue function satisfies the properties

R.1 *Nonnegativity*: $r(\mathbf{p},\mathbf{x}) \geq 0$.

R.2 *Nondecreasing in* **p**: if $\mathbf{p}^0 \geq \mathbf{p}^1$ then $r(\mathbf{p}^0,\mathbf{x}) \geq r(\mathbf{p}^1,\mathbf{x})$.

R.3 *Nondecreasing in* **x**: if $\mathbf{x}^0 \geq \mathbf{x}^1$ then $r(\mathbf{p},\mathbf{x}^0) \geq r(\mathbf{p},\mathbf{x}^1)$.

R.4 *Convex in* **p**: $r(\theta\mathbf{p}^0 + (1-\theta)\mathbf{p}^1,\mathbf{x}) \leq \theta r(\mathbf{p}^0,\mathbf{x}) + (1-\theta) r(\mathbf{p}^1,\mathbf{x})$ for all $0 \leq \theta \leq 1$.

R.5 *Homogeneity*: $r(k\mathbf{p}, \mathbf{x}) = kr(\mathbf{p}, \mathbf{x})$ for $k > 0$.

These properties are analogous to the cost function properties C.1 to C.5 and should be self-explanatory.

As an example, consider the problem of maximising revenue subject to the familiar technology constraint, $q = 2x_1^{0.5}x_2^{0.4}$. In this case, the revenue maximisation problem is

$$r(p, x_1, x_2) = \max_{q} \ pq \quad \text{such that} \quad q - 2x_1^{0.5}x_2^{0.4} = 0. \tag{2.62}$$

Because there is only one output, the technology constraint defines the short-run conditional output supply function[20]:

$$q(p, x_1, x_2) = 2x_1^{0.5}x_2^{0.4} \tag{2.63}$$

Thus, the revenue function is

$$r(p, x_1, x_2) = pq(p, x_1, x_2) = 2px_1^{0.5}x_2^{0.4} \tag{2.64}$$

It is clear from the form of this function that it is nondecreasing in prices and nondecreasing in input quantities.

Because the revenue maximisation and cost minimisation problems are conceptually so similar, it should come as no surprise that we can work backwards from the revenue function to the conditional output supply functions by simply differentiating with respect to output(s). We can also define *short-run revenue functions* by assuming one or more outputs are fixed. Finally, in the single-output case we can define:

Long-run total revenue:	LTR = pq	(2.65)
Long-run average revenue:	LAR = p	(2.66)
Long-run marginal revenue:	LMR = p	(2.67)

When we plot LTR, LAR and LMR against q, the equation for LTR is the equation of a straight line that passes through the origin and has slope p, while the equation for LAR = LMR is a horizontal line with intercept p (see Figures 2.9 and 2.10 below).

2.6 Profit Functions

Until now we have been looking at how firms use input and output price information to choose levels of either inputs or outputs, but not both. In this section we look at how firms choose inputs and outputs simultaneously. We usually assume that firms make these decisions in order to maximise profit (i.e., revenue minus cost). Specifically, we assume multiple-input multiple-output firms solve the problem

$$\pi(\mathbf{p}, \mathbf{w}) = \max_{\mathbf{q},\mathbf{x}} \ \mathbf{p}'\mathbf{q} - \mathbf{w}'\mathbf{x} \quad \text{such that} \quad T(\mathbf{q}, \mathbf{x}) = 0. \tag{2.68}$$

Again, we have used the notation $\pi(\mathbf{p}, \mathbf{w})$ on the left-hand-side to emphasise that maximum profit varies with $\mathbf{p}$ and $\mathbf{w}$.

[20] We use the notation $q(p, x_1, x_2)$ on the left-hand side of 2.63, even though q is not, in this example, a function of p (this is because there is only one output and, consequently, no output-substitution possibilities). When there is more than one output, the conditional output supplies will usually be functions of output prices.

2.6.1 Two Examples

To illustrate the solution to the profit maximisation problem and some associated ideas, consider the problem of maximising profits when the production function takes the simple Cobb-Douglas form $q = x^{0.5}$. In this case, problem 2.68 becomes

$$\pi(p,w) = \max_{q,x} \; pq - wx \quad \text{such that} \quad q - x^{0.5} = 0, \tag{2.69}$$

or, by substituting for q,

$$\pi(p,w) = \max_{x} \; px^{0.5} - wx. \tag{2.70}$$

The first-order condition,

$$0.5px^{-0.5} - w = 0, \tag{2.71}$$

can be solved for the (unconditional) *input demand function*

$$x(p,w) = 0.25w^{-2}p^2 \tag{2.72}$$

Substituting this result back into the production function yields the *output supply function,*

$$q(p,w) = 0.5w^{-1}p, \tag{2.73}$$

and eventually the profit function:

$$\pi(p,w) = pq(p,w) - wx(p,w) = 0.25w^{-1}p^2. \tag{2.74}$$

To make these ideas more concrete, consider the profit-maximising input and output levels when $p = 1$ and $w = 4$. Substituting these prices into equations 2.72 to 2.74 we find $(x, q, \pi) = (0.016, 0.125, 0.063)$. Repeating the exercise using the higher output price $p = 3$ (but still using the input price $w = 4$) we find $(x, q, \pi) = (0.141, 0.375, 0.563)$. That profit should increase when we increase the output price is but one of several plausible properties of the profit function (see Section 2.6.2 below).

We chose this two-variable example because it can be easily represented graphically. To do this we first rewrite the profit function $\pi = pq - wx$ in the form $q = (\pi / p) + (w / p)x$. This equation is the equation of an *isoprofit line* – a straight line with intercept π / p and slope w / p that gives all input-output pairs capable of producing profit level π. In Figure 2.8 we depict the production function $q = x^{0.75}$ and two such isoprofit lines. The isoprofit line having slope $w / p = 4$ is tangent to the production function at the point $(x, q) = (0.016, 0.125)$ and intersects the vertical axis at $\pi / p = 0.063$. The isoprofit line with slope $w / p = 4/3 = 1.33'$ is tangent to

the production function at the point $(x, q) = (0.141, 0.375)$ and intersects the quantity axis at $\pi / p = 0.188$ (and since $p = 3$, this implies $\pi = 0.563$). These are the two profit-maximising solutions computed above.

To help understand the relationship between profit, cost and revenue functions, it will also be useful to solve the profit maximisation problem using the production function $q = 2x_1^{0.5}x_2^{0.4}$ (i.e., the production function used in Sections 2.4 and 2.5). The problem is

$$\pi(p, w_1, w_2) = \max_{q, x_1, x_2} \; pq - (w_1x_1 + w_2x_2) \quad \text{such that} \quad q - 2x_1^{0.5}x_2^{0.4} = 0 \tag{2.75}$$

or, by substituting for q,

$$\pi(p, w_1, w_2) = \max_{x_1, x_2} \; 2px_1^{0.5}x_2^{0.4} - (w_1x_1 + w_2x_2). \tag{2.76}$$

The first-order conditions for a maximum are

$$px_1^{-0.5}x_2^{0.4} - w_1 = 0 \tag{2.77}$$

and $$0.8px_1^{0.5}x_2^{-0.6} - w_2 = 0. \tag{2.78}$$

These first-order conditions can be solved for the *input demand functions*

$$x_1(p, w_1, w_2) = 0.4096w_1^{-6}w_2^{-4}p^{10} \tag{2.79}$$

and $$x_2(p, w_1, w_2) = 0.3277w_1^{-5}w_2^{-5}p^{10}. \tag{2.80}$$

Substituting these input demand functions back into the production function yields

$$q(p, w_1, w_2) = 0.8192w_1^{-5}w_2^{-4}p^{9}. \tag{2.81}$$

Thus, the profit function is

$$\begin{aligned}\pi(p, w_1, w_2) &= pq(p, w_1, w_2) - w_1x_1(p, w_1, w_2) - w_2x_2(p, w_1, w_2) \\ &= 0.0819w_1^{-5}w_2^{-4}p^{10}.\end{aligned} \tag{2.82}$$

Given any set of prices, we can use equations 2.79 to 2.82 to find the profit-maximising input-output combination, and the maximum profit. For example, at prices $(w_1, w_2, p) = (150, 1, 20)$ we find that the profit maximising input-output combination is $(x_1, x_2, q) = (0.368, 44.187, 5.523)$. This yields a maximum profit of $\pi = 11.05$. Interestingly, using these same prices, the cost function 2.29 can be used to show that the minimum cost of producing (the profit-maximising) $q = 5.523$ is $c = 99.42$. Moreover, the revenue function 2.64 tells us that the maximum revenue obtainable from (the profit-maximising) inputs $(x_1, x_2) = (0.368, 44.187)$ is $r = 110.47$. Thus,

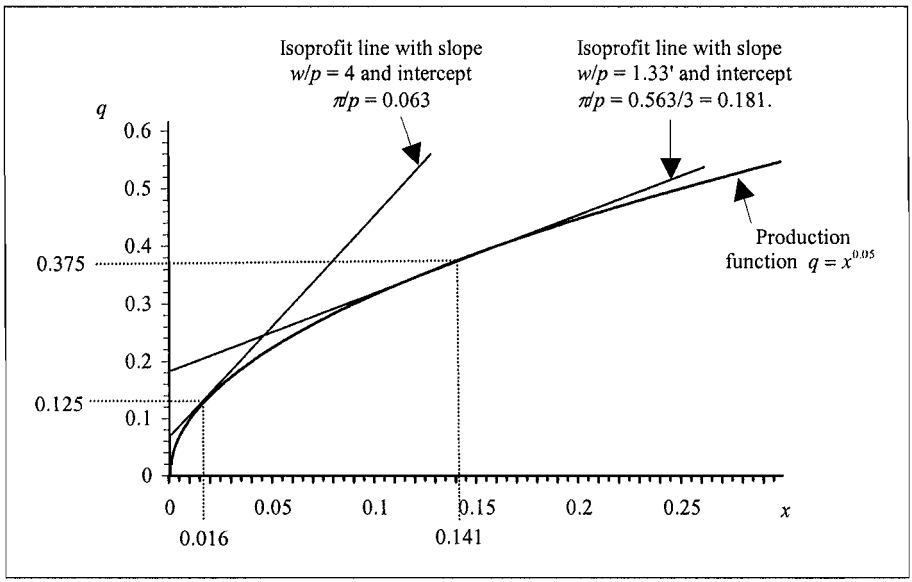

Figure 2.8 Profit Maximisation

$$\begin{aligned} \text{maximum profit} &= \text{maximum revenue} - \text{minimum cost} \\ &= 110.47 - 99.42 \\ &= 11.05 \end{aligned} \tag{2.83}$$

By implication, a firm that maximises profit also maximises revenue and minimises cost.

Finally, we can depict some characteristics of this profit-maximising solution graphically. Figure 2.9 reproduces the LTC function depicted earlier in Figure 2.7 (but it is now drawn on a different scale) along with the LTR function 2.65. The vertical distance between these two functions measures profit, and this is maximised when $q = 5.523$.

2.6.2 Properties

Irrespective of the properties of the underlying transformation function, the profit function will satisfy the following properties:

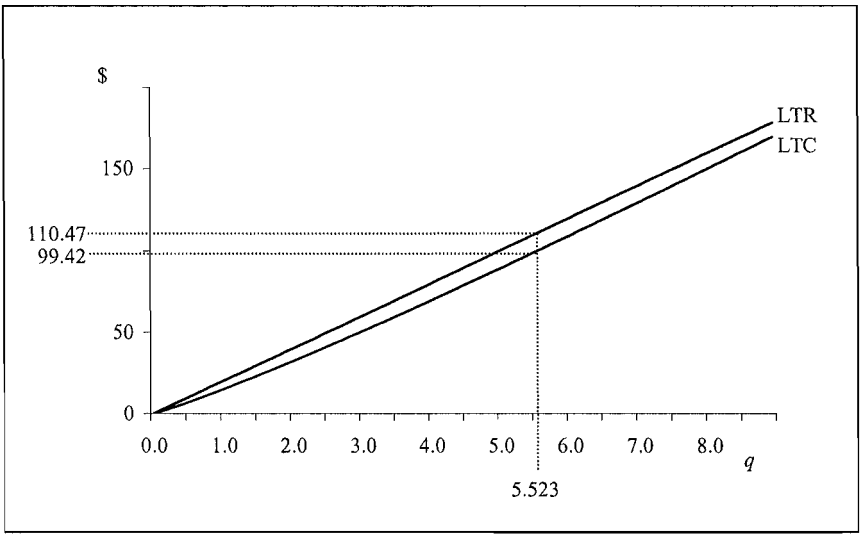

Figure 2.9 LTR, LTC and Profit Maximisation

P.1	*Nonnegativity:*	$\pi(\mathbf{p},\mathbf{w}) \geq 0$
P.2	*Nondecreasing in* $\mathbf{p}$:	if $\mathbf{p}^0 \geq \mathbf{p}^1$ then $\pi(\mathbf{p}^0,\mathbf{w}) \geq \pi(\mathbf{p}^1,\mathbf{w})$.
P.3	*Nonincreasing in* $\mathbf{w}$:	if $\mathbf{w}^0 \geq \mathbf{w}^1$ then $\pi(\mathbf{p},\mathbf{w}^0) \leq \pi(\mathbf{p},\mathbf{w}^1)$.
P.4	*Homogeneity:*	$\pi(k\mathbf{p},k\mathbf{w}) = k\pi(\mathbf{p},\mathbf{w})$ for $k > 0$.
P.5	*Convex in* $(\mathbf{p}, \mathbf{w})$:	$\pi(\theta\mathbf{p}^0 + (1-\theta)\mathbf{p}^1, \theta\mathbf{w}^0 + (1-\theta)\mathbf{w}^1) \leq \theta\pi(\mathbf{p}^1,\mathbf{w}^1) + (1-\theta)\pi(\mathbf{p}^1,\mathbf{w}^1)$ for all $0 \leq \theta \leq 1$.

These properties are generalisations of the properties of cost and revenue functions listed in Sections 2.4.2 and 2.5. Other interesting properties of the profit function, or, more precisely, the profit-maximising solution, can be derived by writing the profit maximisation problem in the form:

$$\pi(\mathbf{p},\mathbf{w}) = \max_{\mathbf{q}} \ \mathbf{p}'\mathbf{q} - c(\mathbf{w},\mathbf{q}). \tag{2.84}$$

In the case of a single-output firm, this implies

$$\pi(p,\mathbf{w}) = \max_{q} \ pq - c(\mathbf{w},q) \tag{2.85}$$

with first-order condition:

$$p - \frac{\partial c(\mathbf{w}, q)}{\partial q} = 0, \tag{2.86}$$

or, LMR = LMC.

Thus, the long-run profit-maximising level of output is the level that equates long-run marginal revenue with long-run marginal cost. In the case of the production function, $q = 2x_1^{0.5}x_2^{0.4}$, the LMC function is given by equation 2.56. Evaluating this function at $(w_1, w_2, p) = (150, 1, 20)$ yields LMC = 20 (= p = LMR). This solution to the profit maximisation problem is depicted in Figure 2.10.

2.6.3 Deriving Input Demand and Output Supply Equations

In Section 2.4.3 we saw how Shephard's Lemma could be used to obtain conditional input demand equations directly from the cost function, without having to explicitly solve an optimisation problem. This idea generalises to the case of a profit function. Specifically, if the profit function is twice-continuously differentiable then *Hotelling's Lemma* says that:

$$x_n(\mathbf{p}, \mathbf{w}) = -\frac{\partial \pi(\mathbf{p}, \mathbf{w})}{\partial w_n} \tag{2.87}$$

and $$q_m(\mathbf{p}, \mathbf{w}) = \frac{\partial \pi(\mathbf{p}, \mathbf{w})}{\partial p_m}. \tag{2.88}$$

To illustrate, consider the profit function 2.82 derived in Section 2.6.1:

$$\pi(p, w_1, w_2) = 0.0819 w_1^{-5} w_2^{-4} p^{10}. \tag{2.82}$$

Applying Hotelling's Lemma:

$$x_1(p, w_1, w_2) = -\frac{\partial \pi(p, w_1, w_2)}{\partial w_1} = 0.4095 w_1^{-6} w_2^{-4} p^{10} \tag{2.89}$$

$$x_2(p, w_1, w_2) = -\frac{\partial \pi(p, w_1, w_2)}{\partial w_2} = 0.3276 w_1^{-5} w_2^{-5} p^{10} \tag{2.90}$$

and $$q(p, w_1, w_2) = \frac{\partial \pi(p, w_1, w_2)}{\partial p} = 0.819 w_1^{-5} w_2^{-4} p^{9}. \tag{2.91}$$

These three equations are identical to the input demand and output supply equations 2.79 to 2.81 (apart from minor rounding errors).

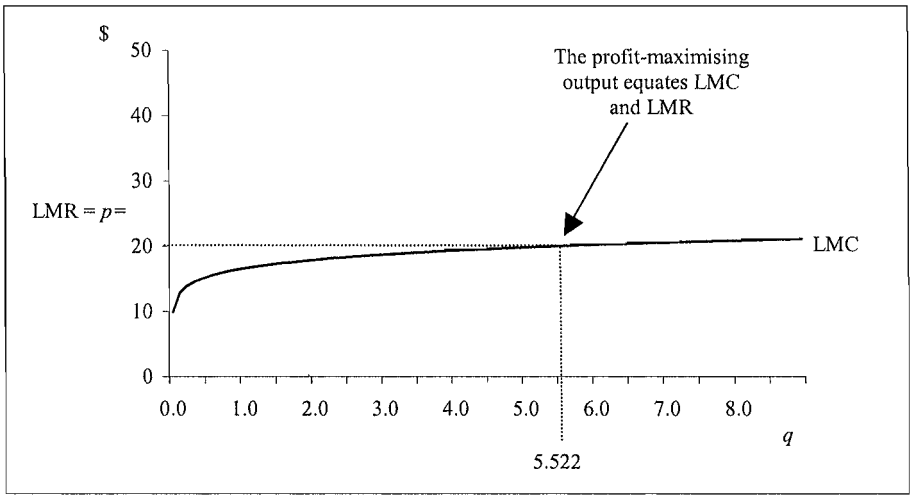

Figure 2.10 LMR, LMC and Profit Maximisation

Finally, if the profit function is twice-continuously differentiable and satisfies properties P.1 to P.5 then Hotelling's Lemma can be used to establish the following properties of input demand and output supply functions:

X.1 *Nonnegativity*: $x_n(\mathbf{p}, \mathbf{w}) \geq 0$.

X.2 *Nonincreasing in* **w**: $\partial x_n(\mathbf{p}, \mathbf{w}) / \partial w_n \leq 0$.

X.3 *Homogeneity*: $x_n(k\mathbf{p}, k\mathbf{w}) = x_n(\mathbf{p}, \mathbf{w})$ for $k > 0$.

X.4 *Symmetry*: $\partial x_n(\mathbf{p}, \mathbf{w}) / \partial w_m = \partial x_m(\mathbf{p}, \mathbf{w}) / \partial w_n$.

Q.1 *Nonnegativity*: $q_m(\mathbf{p}, \mathbf{w}) \geq 0$.

Q.2 *Nondecreasing in* **p**: $\partial q_m(\mathbf{p}, \mathbf{w}) / \partial p_m \geq 0$.

Q.3 *Homogeneity*: $q_m(k\mathbf{p}, k\mathbf{w}) = q_m(\mathbf{p}, \mathbf{w})$ for $k > 0$.

Q.4 *Symmetry*: $\partial q_n(\mathbf{p}, \mathbf{w}) / \partial p_m = \partial q_m(\mathbf{p}, \mathbf{w}) / \partial p_n$.

The nonnegativity properties X.1 and Q.1 stem from the fact that that the profit function is nondecreasing in **p** and nonincreasing in **w** (properties P.2 and P.3). The monotonicity properties X.2 and Q.2 stem from the convexity property of the profit function (property P.5). The homogeneity properties X.3 and Q.3 stem from the fact that the profit function is homogeneous of degree 1 (property P.4) and the result that the first derivative of any linearly homogeneous function is homogeneous of degree 0. Finally, the symmetry properties X.4 and Q.4 follow from the fact that the

order of differentiation is unimportant (Young's Theorem). This means, for example,

$$\frac{\partial x_n(\mathbf{p},\mathbf{w})}{\partial w_m} = \frac{\partial^2 \pi(\mathbf{p},\mathbf{w})}{\partial w_n \partial w_m} = \frac{\partial^2 \pi(\mathbf{p},\mathbf{w})}{\partial w_m \partial w_n} = \frac{\partial x_m(\mathbf{p},\mathbf{w})}{\partial w_n}. \tag{2.92}$$

Again, these properties can be used to explore the effects of possible changes in economic policy. In this book we are just as interested in using them to obtain better estimates of the parameters of input demand and output supply functions (and hence better estimates of economic quantities of interest, including measures of productivity and efficiency).

2.6.4 The Restricted Profit Function

Until now we have assumed that all inputs and outputs in the profit maximisation problem are variable. The profit function 2.68, which treats all inputs and outputs as variable, is sometimes known as an *unrestricted* or *long-run profit function*. Special cases of this function are obtained by assuming that one or more inputs or outputs are fixed, as they would be in the short run. The resulting profit function is known as a *restricted* or *short-run profit function*. We have already considered two restricted profit functions in this chapter: the cost function is (the negative of) a restricted profit function corresponding to the case where all outputs are fixed; and the revenue function is a restricted profit function where all inputs are fixed.

Another restricted profit function is obtained by assuming only a *subset* of inputs are fixed. The resulting short-run profit maximisation problem can be written

$$\pi(\mathbf{p},\mathbf{w},\mathbf{x}_f) = \max_{\mathbf{q},\mathbf{x}_v} \ \mathbf{p}'\mathbf{q} - \mathbf{w}'\mathbf{x} \quad \text{such that} \quad T(\mathbf{q},\mathbf{x}) = 0. \tag{2.93}$$

This problem is identical to the long-run profit maximisation problem 2.68 except we now only search over values of the outputs and *variable* inputs (i.e., $\mathbf{q}$ and $\mathbf{x}_v$). Because we no longer search over (potentially more profitable) values of the fixed inputs, it is clear that short-run profit can never be greater than long-run profit.

The short-run profit function 2.93 satisfies properties P.2 to P.5 (it no longer satisfies the nonnegativity property P.1 because of fixed input costs). In addition, it can be shown to be nonincreasing in fixed inputs. Thai is, if $\mathbf{x}_f^0 \geq \mathbf{x}_f^1$ then $\pi(\mathbf{p},\mathbf{w},\mathbf{x}_f^0) \leq \pi(\mathbf{p},\mathbf{w},\mathbf{x}_f^1)$. It is possible to illustrate these properties using numerical examples of the type used elsewhere in this chapter. However, to avoid repetition we have chosen not to do so here.

2.7 Conclusions

In this chapter we have seen how cost, revenue and profit functions can be derived from production (or transformation) functions by solving constrained optimisation problems. We have also seen how to work back from cost, revenue and profit functions to find input demand and output supply equations (eg., using Hotelling's Lemma). But can we work all the way back to the production technology? The answer is "yes", for reasons that are beyond the scope of this book[21]. However, the very fact that it can be done means the cost, revenue and profit functions must contain essentially the same information as the transformation (or production) function. In fact, it can be shown that every property of a transformation function can be translated into a property of cost, revenue and profit functions, and vice versa. This relationship is known as the principle of *duality*.

An important practical implication of duality theory is that we can use several different types of function to represent all the economically-relevant characteristics of a production technology. Thus, we have the flexibility to choose a representation of the technology that suits our data and our assumptions about the optimising behaviour of firms. For example, we might estimate profit functions when price (and profit) data are available, but we might prefer to estimate cost functions for regulated industries when the profit maximisation assumption is inappropriate. Duality also means, for example, that we can completely specify the production technology by specifying a continuous, monotonic, concave, homogeneous function of prices. This can be very convenient for econometric modeling work.

Unfortunately, the results and methods presented in this chapter are not always applicable. Three situations come to mind. First, the transformation function may not be smooth and continuous, implying we can no longer use results derived using differential calculus. The right-angled isoquant depicted in panel a) of Figure 2.4 is a production technology of this type. This does not prevent us from deriving cost, revenue and profit functions, but it prevents us from doing so using differential calculus.[22] Second, even when we can use differential calculus, we sometimes run into difficulties when the calculus conditions yield multiple optima. Thus, we are unable to identify a unique input-output combination for the firm. Finally, all of the results presented in this chapter were derived under the assumption that the firm is efficient – we assumed the firm knows how to obtain maximum outputs from given inputs, and how to choose input and output mixes to maximise revenues and minimise costs. We will relax this efficiency assumption in Chapter 3.

---

[21] For details see, for example, Chambers (1988, p.281-284).

[22] The right-angled technology depicted in Figure 2.4 is known as a Leontief technology. Varian (1992, p. 54) derives the cost function for a two-input Leontief production function.

3. PRODUCTIVITY AND EFFICIENCY MEASUREMENT CONCEPTS

3.1 Introduction

The production economics concepts discussed in Chapter 2 provide sufficient background for the basic efficiency and productivity measurement methods discussed in this book. Much of the material of Chapter 2 is similar to what is encountered in an undergraduate microeconomics course. The present chapter reviews some additional, more advanced, production economics material. Our primary focus in this chapter is on providing an introduction to a set-theoretic representation of a production technology, thus offering the reader a notional understanding of the framework underlying the concept of the distance function. Distance functions play a crucial role in productivity measurement. The chapter also briefly refers to the cost, revenue and profit functions, introduced in Chapter 2, that are commonly used in studying firms with multiple outputs and inputs. The basic measures of technical efficiency, cost efficiency, allocative efficiency and scale efficiency are defined and their inter-relationships are briefly discussed. The essential difference between the material presented in this chapter and that of Chapter 2 is that we make use of set-theoretic concepts in contrast to the use of functions to describe the production technology in its primal and dual forms.

This chapter is organised as follows. In Section 3.2, we start with a set-theoretic representation of a production technology. The important concepts of multi-output and multi-input distance functions are described in Section 3.3. Various measures of efficiency are introduced and their inter-relationships are examined in detail in Section 3.4. The chapter also introduces the idea of a productivity index, with

particular attention to the Malmquist productivity index, which features prominently throughout the text. Section 3.5 describes various approaches to the measurement of total factor productivity growth. These concepts are further discussed in Chapter 11, which explains how different methods, such as the data envelopment analysis, stochastic frontier models and index numbers, can be used in obtaining estimates of the desired productivity indices discussed in this chapter. The chapter ends with a few concluding remarks in Section 3.6.

3.2 Set Theoretic Representation of a Production Technology

In the previous chapter, we commented that one could easily generalise the single-output cost and profit functions to accommodate multiple-output situations. To avoid potential confusion, we reserve the term *production function* for the case of a single-output technology. Hence, we refer to a multiple-output production process as a multiple-output *production technology* (not to a production function).

A convenient way to describe a multi-input, multi-output production technology is to use the technology set, S. Following Färe and Primont (1995), we use the notation $\mathbf{x}$ and $\mathbf{q}$ to denote a $N\times 1$ input vector of non-negative real numbers and a non-negative $M\times 1$ output vector, respectively.[1] The elements of these vectors are non-negative real numbers. The technology set is then defined as:

$$S = \{(\mathbf{x},\mathbf{q}): \mathbf{x} \text{ can produce } \mathbf{q}\}. \tag{3.1}$$

This set consists of all input-output vectors $(\mathbf{x},\mathbf{q})$, such that $\mathbf{x}$ can produce $\mathbf{q}$. This technology can also be represented using a technical transformation function (see Chapter 2).

The production technology can equivalently be represented and described using output and input sets.

3.2.1 Output Sets

The production technology defined by the set, S, may be equivalently defined using the output set, $P(\mathbf{x})$, which represents the set of all output vectors, $\mathbf{q}$, that can be produced using the input vector, $\mathbf{x}$. Notationally, the output set is defined by

$$P(\mathbf{x}) = \{\mathbf{q}: \mathbf{x} \text{ can produce } \mathbf{q}\} = \{\mathbf{q} : (\mathbf{x},\mathbf{q}) \in S\}. \tag{3.2}$$

The output set is the basis for drawing the production possibility curves for two-dimensional output vectors commonly seen in textbooks. The output sets are

[1] Thus, $\mathbf{x}$ and $\mathbf{q}$ can be considered as elements of the non-negative orthant of N- and M-dimensional Euclidean spaces, represented by R_N^+ and R_M^+, respectively.

sometimes referred to as production possibility sets associated with various input vectors, **x**.

The properties of the output set can be summarised as follows. For each **x**, the output set P(**x**) is assumed to satisfy:

(i) $\mathbf{0} \in P(\mathbf{x})$: nothing can be produced from a given set of inputs (i.e., inaction is possible);

(ii) non-zero output levels cannot be produced from zero levels of inputs;

(iii) P(**x**) satisfies strong disposability of outputs: if $\mathbf{q} \in P(\mathbf{x})$ and $\mathbf{q}^* \leq \mathbf{q}$ then $\mathbf{q}^* \in P(\mathbf{x})$;[2]

(iv) P(**x**) satisfies strong disposability of inputs: if **q** can be produced from **x**, then **q** can be produced from any $\mathbf{x}^* \geq \mathbf{x}$;[3]

(v) P(**x**) is closed;

(vi) P(**x**) is bounded; and

(vii) P(**x**) is convex.

The assumption of closedness is essentially a mathematical requirement[4], but the bounded nature of P(**x**) implies that we cannot produce unlimited levels of outputs with a given set of inputs. Convexity implies that if two combinations of output levels can be produced with a given input vector **x**, then any weighted average of these output vectors can also be produced. This assumption implicitly requires that the commodities are continuously divisible.

3.2.2 Input Sets

The input associated with a given output vector, **y**, is defined as the set:

$$L(\mathbf{q}) = \{\mathbf{x}: \mathbf{x} \text{ can produce } \mathbf{q}\} = \{\mathbf{x}: (\mathbf{x}, \mathbf{q}) \in S\}. \tag{3.3}$$

The input set consists of all input vectors, **x**, that can produce a given output vector, **q**. Given the basic assumptions on the production technology, the following properties of the input sets can be derived.

(i) L(**q**) is closed for all **q**;

---

[2] An alternative assumption to strong disposability is "weak disposability", which states that if a vector of outputs, **q**, can be produced from a given input vector, **x**, then any contraction of **q**, $\lambda\mathbf{q}$, with $0<\lambda<1$, can also be produced with **x**. It is easy to see that strong disposability implies weak disposability, but not vice versa.

[3] Since $\mathbf{x}^*$ and **x** are vectors, then $\mathbf{x}^* \geq \mathbf{x}$ holds when all elements of $\mathbf{x}^*$ are greater than or equal to the corresponding elements in **x**, but strictly greater for at least one element.

[4] See Färe and Primont (1995), p.14, for further discussion on this property.

(ii) L(**q**) is convex for all **q**;

(iii) Inputs are said to be weakly disposable if $\mathbf{x} \in L(\mathbf{q})$ then, for all $\lambda \geq 1$, $\lambda\mathbf{x} \in L(\mathbf{q})$; and

(iv) Inputs are said to be strongly disposable if $\mathbf{x} \in L(\mathbf{q})$ and if $\mathbf{x}^* \geq \mathbf{x}$ then $\mathbf{x}^* \in L(\mathbf{q})$.

These properties of the input distance function can be easily derived using the assumptions made with respect to the production technology implicit in the properties of P(**x**).

As the output and input sets provide alternative descriptions of the same underlying technology, these two sets are also interrelated. It can be easily seen that if **q** belongs to P(**x**), i.e., **q** can be produced using input vector **x**, then **x** belongs to the input set of **q**, L(**q**). It is important to realise that these descriptions are equivalent because they contain the same information.

In this section, we have not explicitly accounted for time. If the production technology under consideration refers to year t, it is necessary to label the sets S, P and L as S^t, $P^t(\mathbf{x})$ and $L^t(\mathbf{x})$. This kind of labelling is very important when productivity levels are measured at two different time points and productivity change over time is of interest. Thus, Section 3.5 uses input, output and technology sets with a time subscript.

Before we use this framework to define multi-input and multi-output distance functions, we briefly digress to discuss the output sets using production possibility curves.

3.2.3 A Digression on Production Possibility Curves and Revenue Maximisation

A multi-output production technology can be very difficult to conceptualise or to visualise. We can attempt to provide some understanding by using a simple one-input, two-output example. In this instance, we specify an input requirement function where we express the single input as a function of the two outputs:

$$x_1 = g(q_1, q_2). \tag{3.4}$$

This one-input, two-output case can be used to illustrate the idea of a *production possibility curve* (PPC), which is the output counterpart of the isoquant. The isoquant represents the various combinations of inputs that could be used to produce a given output level. The production possibility curve, on the other hand, depicts the various output combinations that could be produced using a given input level. An example of a production possibility curve is provided in Figure 3.1.

A discussion of the properties of this curve would follow similar lines to our discussion of the isoquant, so we omit much of it. Obviously a production

possibility curve could be drawn for each input level. Furthermore, we observe that the combination of outputs that maximise profit, given an input level, are equivalent to that which maximises revenue.[5] The revenue equivalent to the isocost line is the *isorevenue line,* which has slope equal to $(-p_1/p_2)$, the negative ratio of the output prices. The optimal (revenue-maximising) point is determined by the point of tangency between this line and the production possibility curve, as depicted in Figure 3.2.

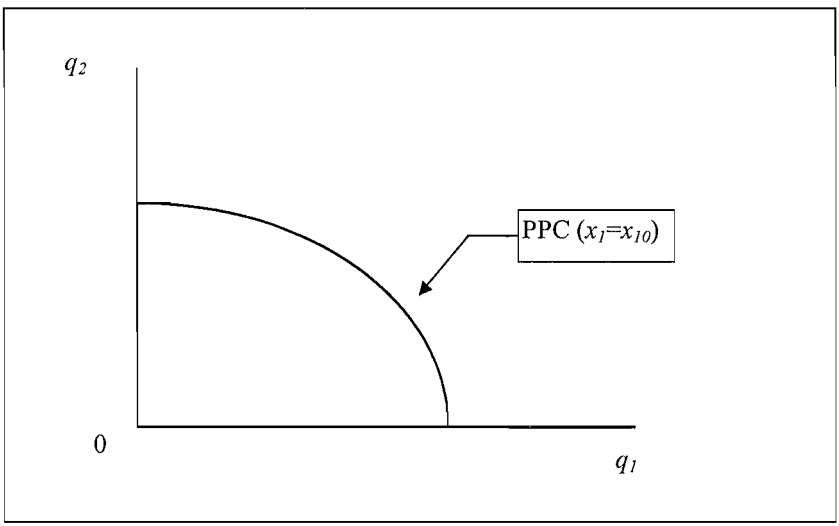

Figure 3.1 Production Possibility Curve

Production at any point on the production possibility curve other than point A in Figure 3.2 coincides with an isorevenue curve which is closer to the origin and, hence, implies a lower total revenue (and, thus, a lower profit).

Before we return to our discussion of production sets and distance functions, we quickly make note of the fact that our discussion of biased technical change in the previous chapter can be extended to include multiple output situations. Technical change can favour the production of one commodity over another. This concept is illustrated in Figure 3.3.

[5] This is similar to the notion that selecting input levels so as to minimise the cost of producing a given output level, is equivalent to maximising profit (given the output constraint).

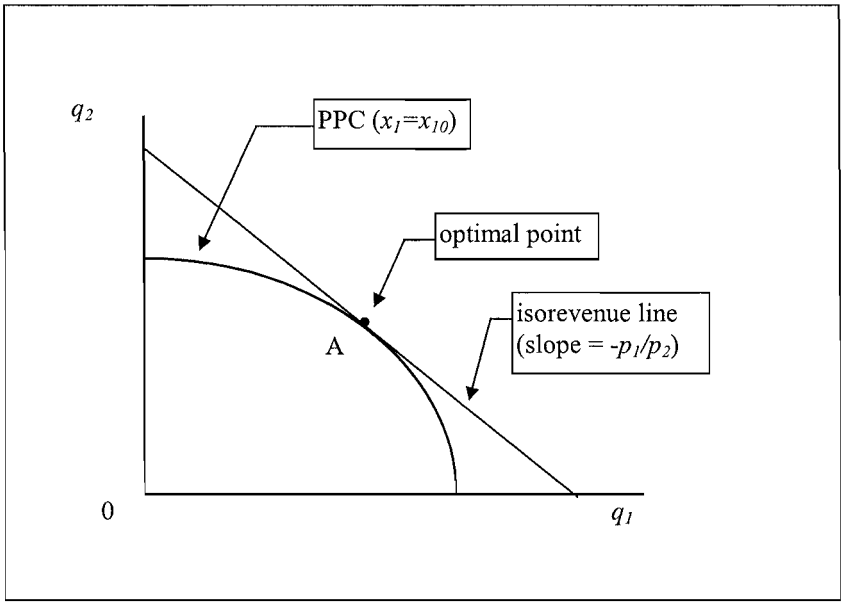

Figure 3.2 The Production Possibility Curve and Revenue Maximisation

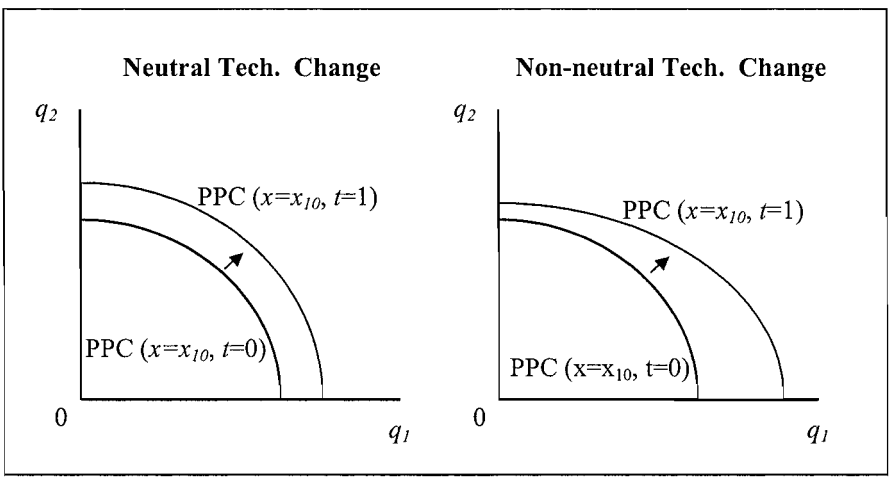

Figure 3.3 Technical Change and the Production Possibility Curve

3.3 Output and Input Distance Functions

Distance functions are very useful in describing the technology in a way that makes it possible to measure efficiency and productivity. The concept of a distance function is closely related to production frontiers. The basic idea underlying distance functions is quite simple, involving radial contractions and expansions in defining these functions. The notion of a distance function was introduced independently by Malmquist (1953) and Shephard (1953), but they have gained prominence only in the last three to four decades. This section provides basic definitions, followed by a description of their use in measuring technical efficiency in Section 3.4.

Distance functions allow one to describe a multi-input, multi-output production technology without the need to specify a behavioural objective (such as cost-minimisation or profit-maximisation). One may specify both input distance functions and output distance functions. An input distance function characterises the production technology by looking at a minimal proportional contraction of the input vector, given an output vector. An output distance function considers a maximal proportional expansion of the output vector, given an input vector. We first consider an output distance function.

3.3.1 Output Distance Functions

The output distance function is defined on the output set, $P(\mathbf{x})$, as:

$$d_o(\mathbf{x},\mathbf{q}) = \min\{\delta\colon (\mathbf{q}/\delta)\in P(\mathbf{x})\}.^{6} \tag{3.5}$$

A few simple properties of $d_o(\mathbf{x},\mathbf{q})$ follow[7] directly from the axioms on the technology set:

(i) $d_o(\mathbf{x},\mathbf{0}) = 0$ for all non-negative $\mathbf{x}$;

(ii) $d_o(\mathbf{x},\mathbf{q})$ is non-decreasing in $\mathbf{q}$ and non-increasing in $\mathbf{x}$;

(iii) $d_o(\mathbf{x},\mathbf{q})$ is linearly homogeneous in $\mathbf{q}$;[8]

(iv) $d_o(\mathbf{x},\mathbf{q})$ is quasi-convex[9] in $\mathbf{x}$ and convex in $\mathbf{q}$;[10]

---

[6] The definition of the output distance function in equation (3.5) is made more rigorous by replacing "min" (which stands for "minimum") with "inf" (which stands for "infimum"). This allows for the possibility that the minimum may not exist (i.e., that $\delta = +\infty$ is possible). We shall, however, continue to use the less precise term "min" in this book in the interests of ease of reading.

[7] See Färe and Primont (1995) for detailed proofs and derivations of these properties.

[8] This property follows from distance function definition and not from the properties of the technology.

[9] A function $f(\mathbf{x})$ defined in a convex set in R_n is said to be quasi-convex if and only if for any pair of distinct points $\mathbf{x}$ and $\mathbf{y}$ in the domain of f and $0 < \lambda < 1$, $f(\mathbf{y}) \geq f(\mathbf{x})$ implies that $f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) \leq f(\mathbf{y})$.

[10] A real valued function $f(\mathbf{x})$ defined in a convex set in R_n is convex on the set if $f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x})+(1-\lambda)f(\mathbf{y})$ for any $\mathbf{x},\mathbf{y}$ belonging to the set and $\lambda \geq 0$.

(v) if **q** belongs to the production possibility set of **x** (i.e., $\mathbf{q} \in P(\mathbf{x})$), then $d_o(\mathbf{x},\mathbf{q}) \leq 1$; and

(vi) distance is equal to unity (i.e., $d_o(\mathbf{x},\mathbf{q}) = 1$) if **q** belongs to the "frontier" of the production possibility set (the PPC curve of **x**).[11]

It is useful to illustrate the concept of an output distance function using an example where two outputs, q_1 and q_2, are produced using the input vector, **x**. For a given input vector, **x**, we can represent the production technology on the two dimensional diagram in Figure 3.4. Here the production possibility set, P(**x**), is the area bounded by the production possibility frontier, PPC-P(**x**), and the q_1 and q_2 axes. The value of the distance function for the firm using input level **x** to produce the outputs, defined by the point A, is equal to the ratio δ=0A/0B.

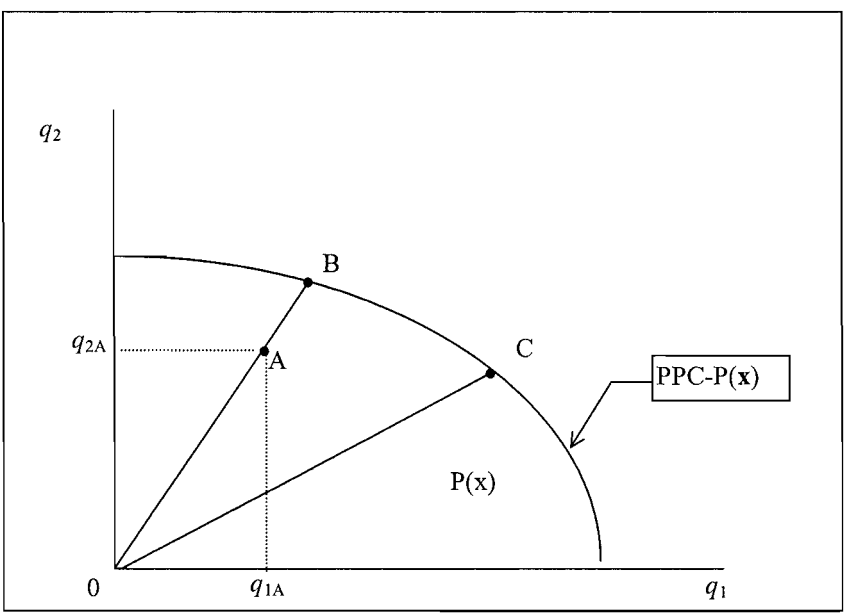

Figure 3.4 Output Distance Function and Production Possibility Set

This distance measure is the reciprocal the factor by which the production of all output quantities could be increased while still remaining within the feasible production possibility set for the given input level. We also observe that the points B and C are on the production possibility surface, denoted by PPC-P(**x**), and, hence, would have distance function values equal to 1.

[11] These properties play a major role in the ensuing chapters which focus on efficiency measurement.

3.3.2 Input Distance Functions

The input distance function, which involves the scaling of the input vector, is defined on the input set, L(**q**), as:

$$d_i(\mathbf{x},\mathbf{q}) = \max\{\rho\colon (\mathbf{x}/\rho)\in \mathrm{L}(\mathbf{q})\},^{12} \tag{3.6}$$

where the input set L(**q**) represents the set of all input vectors, **x**, which can produce the output vector, **q**.

Given the general axioms listed in the Section 3.2, it is easy to show[13] that:

(i) the input distance function is non-decreasing in **x** and non-increasing in **q**;

(ii) it is linearly homogeneous in **x**;[14]

(iii) $d_i(\mathbf{x},\mathbf{q})$ is concave in **x** and quasi-concave in **q**;[15]

(iv) if **x** belongs to the input set of **q** (i.e., $\mathbf{x}\in \mathrm{L}(\mathbf{q})$) then $d_i(\mathbf{x},\mathbf{q}) \geq 1$; and

(v) distance is equal to unity (i.e., $d_i(\mathbf{x},\mathbf{q}) = 1$) if **x** belongs to the "frontier" of the input set (the isoquant of **q**).

We illustrate the input distance function using an example where two inputs, x_1 and x_2, are used in producing output vector, **q**. Now for a given output vector we can represent the production technology on the two-dimensional diagram in Figure 3.5. Here the input set, L(**q**), is the area bounded from below by the isoquant, Isoq-L(**q**). The value of the distance function for the point, A, which defines the production point where firm A uses x_{1A} of input 1 and x_{2A} of input 2, to produce the output vector **q**, is equal to the ratio ρ=0A/0B.

---

[12] Note that the definition of the input distance function in equation (3.6) could be made more rigorous by replacing "max" (which stands for "maximum") with "sup" (which stands for "supremum"). This allows for the possibility that the maximum does not exist (i.e., that $\rho = +\infty$ is possible). We shall, however, continue to use the less precise term "max" in this book in the interests of ease of reading.

[13] See Färe and Primont (1995) for derivations.

[14] This property follows from the definition of input distance function and not on the properties of the technology.

[15] A function $f(\mathbf{x})$ is quasi-concave if and only if $-f(\mathbf{x})$ is quasi-convex. Similarly $f(\mathbf{x})$ is concave if and only $-f(\mathbf{x})$ is convex.

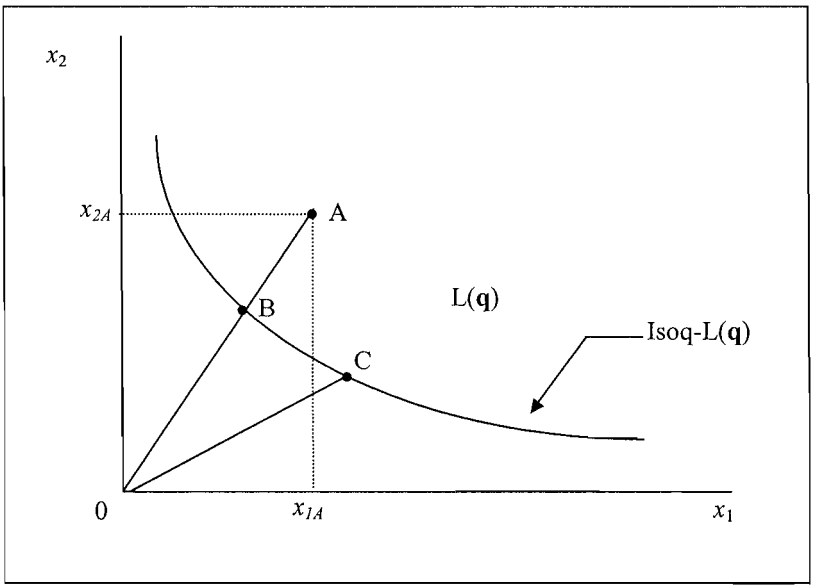

Figure 3.5 Input Distance Function and Input Requirement Set

It is useful to state a couple of results that connect the input and output distance functions. The first point is that if $\mathbf{q} \in P(\mathbf{x})$, then $\mathbf{x} \in L(\mathbf{q})$, i.e., if $\mathbf{q}$ belongs to the production possibility set associated with input vector $\mathbf{x}$, then $\mathbf{x}$ belongs to the feasible input set associated with output vector $\mathbf{q}$.

If both inputs and outputs are weakly disposable, we can state that

$$d_i(\mathbf{x},\mathbf{q}) \geq 1 \text{ if and only if } d_o(\mathbf{x},\mathbf{q}) \leq 1.$$

Further, if the technology exhibits global constant returns to scale then we can state that:

$$d_i(\mathbf{x},\mathbf{q}) = 1/d_o(\mathbf{x},\mathbf{q}), \text{ for all } \mathbf{x} \text{ and } \mathbf{q}.$$

This means that, under constant returns to scale[16], the input distance function is the reciprocal of the output distance function, for any $(\mathbf{x}, \mathbf{q})$.

For further results and discussion, the reader is referred to the excellent, but highly mathematical, treatment of distance functions in Färe and Primont (1995), Balk (1998) and Färe, Grosskopf and Russell (1998).

[16] In fact this condition is necessary and sufficient for this relationship between the input and output distance functions to hold.

Output and input distance functions have a number of applications. They are used in defining a variety of index numbers, as illustrated in Chapter 4. They also provide the conceptual underpinning for various efficiency and productivity measures. These distance functions can be directly estimated using either econometric or mathematical programming methods. Data envelopment analysis (DEA), described in Chapters 6 and 7, is a non-stochastic non-parametric method for identifying production frontiers and for computing input and output distances. Coelli and Perelman (1999, 2000) and O'Donnell and Coelli (2004) provide a discussion of methods for estimating parametric stochastic frontier specification of the distance functions and illustrate their use in an analysis of technical efficiency in European railways. Estimated distance functions have also been used in deriving measures of shadow prices (Färe *et al.* 1989, 1993). Direct econometric estimation of distance function is described in Section 10.2.

3.4 Efficiency Measurement using Distance, Cost and Revenue Functions

The primary purpose of this section is to outline a number of commonly-used efficiency measures and to discuss how they may be calculated relative to a given technology that is represented by some form of frontier function. This section defines various measures of efficiency and describes their relationship with some of the concepts developed thus far. In particular, the concepts of output and input distance functions are discussed in Section 3.3 and the cost and revenue functions are discussed in Chapter 2.

The discussion in this section also provides a very brief introduction to modern efficiency measurement. A more detailed treatment is provided by Färe, Grosskopf and Lovell (1985, 1994) and Lovell (1993). Our discussion of efficiency measurement begins with Farrell (1957), who drew upon the work of Debreu (1951) and Koopmans (1951) to define a simple measure of firm efficiency that could account for multiple inputs. Farrell (1957) proposed that the efficiency of a firm consists of two components: *technical efficiency*, which reflects the ability of a firm to obtain maximal output from a given set of inputs, and *allocative efficiency*, which reflects the ability of a firm to use the inputs in optimal proportions, given their respective prices and the production technology. These two measures are then combined to provide a measure of total *economic efficiency*.[17]

The following discussion begins with Farrell's original ideas that were illustrated in input/input space and, hence, had an input-reducing focus. These are usually termed *input-orientated* measures.

[17] Some of Farrell's terminology differs from that used here. He used the term *price efficiency* instead of *allocative efficiency* and the term *overall efficiency* instead of *economic efficiency*. The terminology used in this book conforms with that which is used most often in recent literature.

3.4.1 Input-Orientated Measures

Farrell illustrated his ideas using a simple example involving firms that use two inputs (x_1 and x_2) to produce a single output (q), under the assumption of constant returns to scale.[18] Knowledge of the unit isoquant of *fully efficient firms*,[19] represented by SS′ in Figure 3.6, permits the measurement of technical efficiency. If a given firm uses quantities of inputs, defined by the point P, to produce a unit of output, the technical inefficiency of that firm could be represented by the distance QP, which is the amount by which all inputs could be proportionally reduced without a reduction in output. This is usually expressed in percentage terms by the ratio QP/0P, which represents the percentage by which all inputs need to be reduced to achieve technically efficient production. The technical efficiency (TE) of a firm is most commonly measured by the ratio

$$TE = 0Q/0P, \tag{3.7}$$

which is equal to one minus QP/0P. It takes a value between zero and one, and, hence, provides an indicator of the degree of technical efficiency of the firm. A value of one implies that the firm is fully technically efficient. For example, the point Q is technically efficient because it lies on the efficient isoquant.

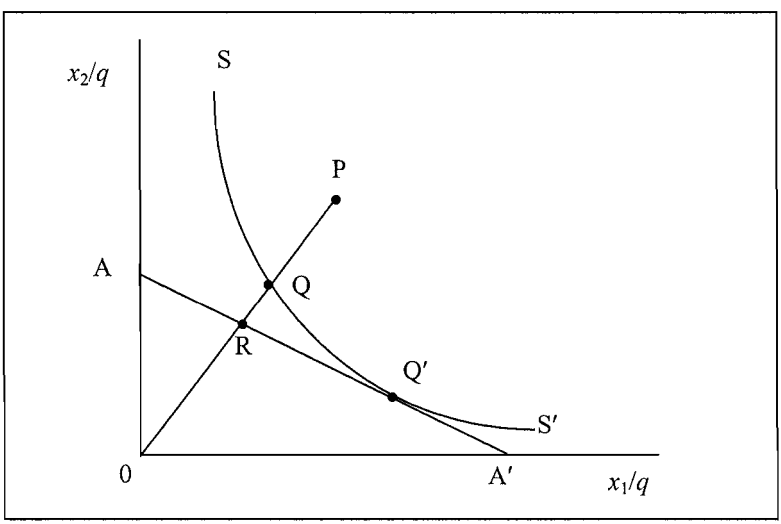

Figure 3.6 Technical and Allocative Efficiencies

[18] The assumption of constant returns to scale allows the technology to be represented using the unit isoquant. Farrell also discussed the extension of his method so as to accommodate more than two inputs, multiple outputs, and non-constant returns to scale.

[19] The production frontier of fully efficient firms is not known in practice, and, thus, must be estimated from observations on a sample of firms in the industry concerned. Estimation of frontiers using DEA and SFA methods are discussed in later chapters.

The input-orientated measure of technical efficiency of a firm can be expressed in terms of input-distance function $d_i(\mathbf{x},\mathbf{q})$ as:

$$TE = 1/d_i(\mathbf{x},\mathbf{q}). \tag{3.8}$$

The firm under consideration is technically efficient if it is on the frontier, in which case $TE = 1$ and $d_i(\mathbf{x},\mathbf{q})$ is also equal to 1.

In the presence of input price information, it would be possible to measure the *cost efficiency* of the firm under consideration. Let $\mathbf{w}$ represent the vector of input prices and let $\mathbf{x}$ represent the observed vector of inputs used associated with point P. Let $\hat{\mathbf{x}}$ and $\mathbf{x}^*$ represent the input vector associated with the technically efficient point Q and the cost-minimising input vector at Q', respectively.

Then cost efficiency of the firm is defined as the ratio of input costs associated with input vectors, $\mathbf{x}$ and $\mathbf{x}^*$, associated with points, P and Q'. Thus

$$CE = \frac{\mathbf{w}'\mathbf{x}^*}{\mathbf{w}'\mathbf{x}} = OR/OP. \tag{3.9}$$

If the input price ratio, represented by the slope of the isocost line, AA', in Figure 3.6, is also known, then allocative efficiency and technical efficiency measures can be calculated using the isocost line. These are given by

$$\begin{aligned} AE &= \frac{\mathbf{w}'\mathbf{x}^*}{\mathbf{w}'\hat{\mathbf{x}}} = \frac{OR}{OQ} \\ TE &= \frac{\mathbf{w}'\hat{\mathbf{x}}}{\mathbf{w}'\mathbf{x}} = \frac{OQ}{OP} \end{aligned} \tag{3.10}$$

These equations follow from the observation that the distance RQ represents the reduction in production costs that would occur if production were to occur at the allocatively (and technically) efficient point Q', instead of at the technically efficient, but allocatively inefficient, point Q.[20]

Given the measure of technical efficiency, the total *overall cost efficiency (CE)* can be expressed as a product of technical and allocative efficiency measures:

$$TE \times AE = (0Q/0P) \times (0R/0Q) = (0R/0P) = CE.$$

[20] One could illustrate this by drawing two isocost lines through Q and Q'. Irrespective of the slope of these two parallel lines (which is determined by the input price ratio) the ratio $RQ/0Q$ represents the proportional reduction in costs of production associated with movement from Q to Q'.

Note also that all three measures are bounded by zero and one[21].

The above graphical illustration of efficiency measures made use of constant-returns-to-scale technology. The use of constant returns to scale and two input variables makes it possible to draw the necessary graphs in two dimensions. These measures can be equivalently defined for the non-constant returns to scale case using simple algebraic expressions. To illustrate this, we could adjust Figure 3.6 by changing the axes labels to x_1 and x_2 and assuming that the isoquant represents the lower bound of the input set associated with the production of a particular level of output. The efficiency measures are then defined analogously to those above.

These efficiency measures assume that the production technology is known. In practice, this is not the case, and the efficient isoquant must be estimated from the sample data. Identifying the production frontier is a complex problem. Chapters 6, 7, 9 and 10 of this book are devoted to this problem of determining frontiers using firm-level data.

3.4.2 Output-Orientated Measures

The above input-orientated technical efficiency measures address the question: "By how much can input quantities be proportionally reduced without changing the output quantities produced?" One could alternatively ask the question: "By how much can output quantities be proportionally expanded without altering the input quantities used?" This gives output-orientated measures, as opposed to the input-orientated measures discussed above. The difference between the output- and input-orientated measures can be illustrated using a simple example involving one input, x, and one output, q. This is depicted in Figure 3.7(a) where we have decreasing-returns-to-scale technology, represented by $f(x)$, and an inefficient firm operating at the point P. The Farrell input-orientated measure of *TE* is equal to the ratio AB/AP, while the output-orientated measure of *TE* is represented by CP/CD. The output- and input-orientated measures are equivalent measures of technical efficiency only when constant returns to scale exist (Färe and Lovell, 1978). The constant-returns-to-scale (CRS) case is depicted in Figure 3.7(b), where we observe that AB/AP=CP/CD, for the inefficient firm operating at point *P*.

[21] This result implies that cost efficiency is always less than or equal to technical efficiency. Thus *CE* is less than or equal to $1/d_i(\mathbf{x},\mathbf{q})$.

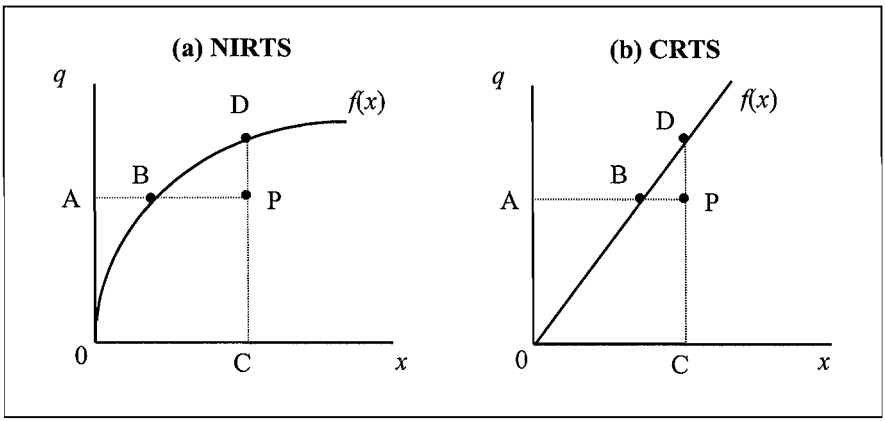

Figure 3.7 Input- and Output-Orientated Technical Efficiency Measures and Returns to Scale

One can illustrate output-orientated measures by considering the case where production involves two outputs (q_1 and q_2) and a single input (x). If we assume CRS we can represent the technology by a unit production possibility curve in two dimensions. This example is depicted in Figure 3.8, where the curve ZZ' is the unit production possibility curve and the point A corresponds to an inefficient firm. Note that an inefficient firm operating at point A lies *below* the curve, because ZZ' represents the upper bound of the production possibilities.

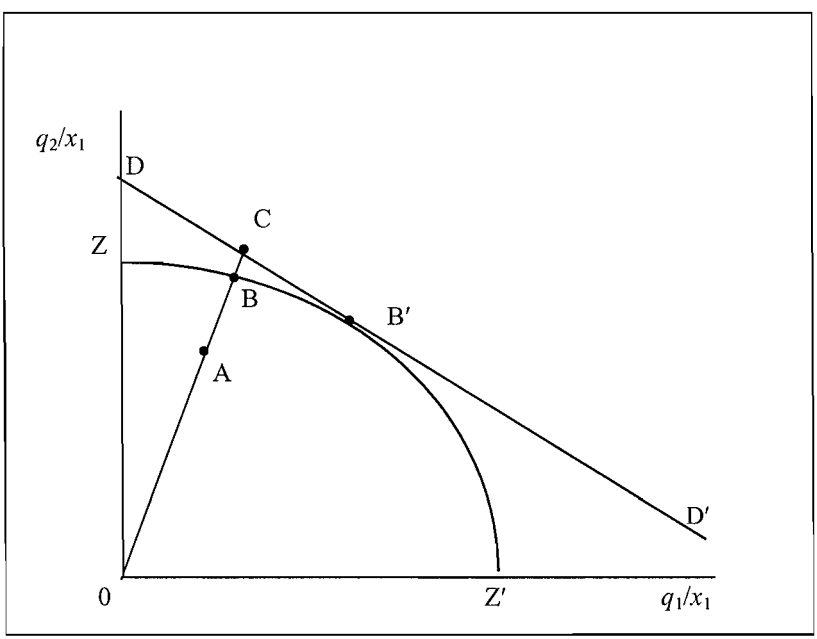

Figure 3.8 Technical and Allocative Efficiencies from an Output Orientation

The Farrell output-orientated efficiency measures (see Färe, Grosskopf and Lovell, 1985, 1994) are defined as follows. In Figure 3.8, the distance AB represents technical inefficiency, which is the amount by which outputs could be increased without requiring extra input. Hence, a measure of output-orientated technical efficiency is the ratio

$$TE = 0A/0B = d_o(\mathbf{x},\mathbf{q}). \tag{3.11}$$

where $d_o(\mathbf{x},\mathbf{q})$ is the output distance function at the observed input vector $\mathbf{x}$ and the observed output vector $\mathbf{q}$.

Now *revenue efficiency* can be defined for any observed output price vector $\mathbf{p}$ represented by the line DD′. If $\mathbf{q}$, $\hat{\mathbf{q}}$ and $\mathbf{q}^*$ represent the observed output vector of firm associated with point A, the technically efficient production vector associated with B and the revenue efficient vector associated with the point B′, respectively, then *revenue efficiency* of the firm is defined as:

$$RE = \frac{\mathbf{p}'\mathbf{q}}{\mathbf{p}'\mathbf{q}^*} = \frac{OA}{OC}. \tag{3.12}$$

If we have price information then we can draw the isorevenue line, DD′, and define the allocative and technical efficiency measures as below:

$$\begin{aligned} AE &= \frac{\mathbf{p}'\hat{\mathbf{q}}}{\mathbf{p}'\mathbf{q}^*} = \frac{OB}{OC} \\ TE &= \frac{\mathbf{p}'\mathbf{q}}{\mathbf{p}'\hat{\mathbf{q}}} = \frac{OA}{OB} \end{aligned} \tag{3.13}$$

which has a revenue-increasing interpretation (similar to the cost-reducing interpretation of allocative inefficiency in the input-orientated case). Furthermore, we define overall revenue efficiency as the product of these two measures

$$RE = (0A/0C) = (0A/0B)\times(0B/0C) = TE\times AE.$$

Again, we note that all of these three measures are bounded by zero and one[22]. We also observe that the output-orientated technical efficiency measure is exactly equal to the output distance function, introduced in Section 3.3.

Before we conclude this section, we note three points about the efficiency measures we have defined. First, technical efficiency has been measured along a ray from the origin to the observed production point. Hence, these measures hold the

[22] This equation implies that *revenue efficiency* is always less than or equal to technical efficiency and that revenue efficiency is greater than or equal to $1/d_o(\mathbf{x},\mathbf{q})$.

relative proportions of inputs (or outputs) constant. One advantage of these *radial* efficiency measures is that they are *units invariant*. That is, changing the units of measurement (e.g., measuring quantity of labour in person hours instead of person years) does not change the value of the efficiency measure. A non-radial measure, such as the shortest distance from the production point to the production surface, seems intuitively appealing, but such a measure is not invariant to the units of measurement. Changing the units of measurement, in this case, could result in the identification of a different "nearest" point.[23]

Second, we have discussed allocative efficiency from a cost-minimising perspective and from a revenue-maximising perspective, but not from a profit-maximising perspective (where both cost minimisation and revenue maximisation are assumed). Profit maximisation can be accommodated in a number of ways. The principal difficulty is associated with the selection of the orientation in which to measure technical efficiency (input, output or both). One suggestion is presented in Färe, Grosskopf and Lovell (1994), in which DEA is used to measure profit efficiency along with a hyperbolic measure of technical efficiency (which considers simultaneous expansion of outputs and contraction of inputs). This requires the use of *directional distance functions* that are technically beyond the scope of this book. This function was introduced by Chambers, Chung and Färe (1996). See Balk (1998) for an illustration of how directional distance functions can be used in dealing with profit efficiency and productivity change. The difference between the two measures is then interpreted as allocative efficiency.[24] An alternative approach is suggested by Kumbhakar (1987) in a stochastic frontier framework, and involves the decomposition of profit efficiency into three components: input-allocative efficiency, output-allocative efficiency and input-orientated technical efficiency. No particular profit efficiency methodology has become widely used to date. The references suggested above provide a reasonable starting point for researchers who wish to explore this issue.

Finally, we repeat our observations that the Farrell input- and output-orientated technical efficiency measures are equivalent to the input and output distance functions, discussed in Shephard (1970) and Färe and Primont (1995).[25] This observation is especially important when we discuss the use of DEA methods in calculating Malmquist indices of TFP change in Chapter 10.

---

[23] A number of alternative non-radial efficiency measures have been proposed which sacrifice units-invariance but have other desirable properties, e.g., see Färe and Lovell (1978) and Kopp (1981). The problem of unit-invariance can be solved if the direction vector is correspondingly changed when units of measurement are changed.

[24] A related suggestion has recently been proposed which involves the use of newly-developed *directional distance functions* which also involve simultaneous expansion of outputs and reduction in inputs. Refer to Färe and Grosskopf and Weber (1997) for more on this method.

[25] In fact, as they are defined in this book, the input-orientated technical efficiency measure is equal to the inverse of the input distance function.

3.4.3 Scale efficiency

So far we have discussed the efficiency of operations of a firm with respect to the production technology frontier and at a given level of input and output prices. It is possible that a firm is both technically and allocatively efficient but the scale of operation of the firm may not be optimal. Suppose the firm is using a variable-returns-to-scale (VRS) technology. Then, the firm involved may be too small in its scale of operation, which might fall within the *increasing returns to scale (irs)* part of the production function. Similarly, a firm may be too large and it may operate within the *decreasing returns to scale* part of the production function. In both of these cases, efficiency of the firms might be improved by changing their scale of operations, i.e., to keep the same input mix but change the size of operations. If the underlying production technology is a globally *constant returns-to-scale (CRS)* technology then the firm is automatically scale efficient.

There have been several attempts to measure scale efficiency and its influence on productivity change over time. Some of the earlier attempts to measure scale efficiency are Førsund and Hjalmarsson (1979, 1987), Banker and Thrall (1992) and Färe, Grosskopf and Lovell (1994). Färe, Grosskopf and Roos (1998) present a definition of scale efficiency and use it in deriving a decomposition of productivity change over time. Balk (2001) provides a formal framework to define scale efficiency and to study the role of scale efficiency in productivity change. Balk then compares and evaluates some of the earlier attempts in the literature (Färe *et. al*, 1994; Ray and Desli, 1997; Grifell-Tatje and Lovell, 1999; Wheelock and Wilson, 1999; and Zofio and Lovell, 1999) to decompose productivity change into efficiency change, technical change and scale change.

Scale efficiency is a simple concept that is easy to understand in a one–input, one–output case, but it is more difficult to conceptualise in a multi-input, multi-output situation. Hence, we first discuss the one-input, one-output case and then provide a brief outline of the multi-input, multi-output case.[26]

A one-input, one-output VRS production technology is depicted in Figure 3.9. The production set, S, is the area between the VRS production frontier, $f(x)$, and the x-axis, inclusive of these bounds. The firms operating at the points A, B and C are all technically efficient, because they are operating on the production frontier. However, because the productivity of each of these firms is equal to the ratio of their observed output and input quantities (i.e., y/x), and this expression is equivalent to the slope of a ray drawn from the origin through the data point (x,y), we can see that even though these three firms are all technically efficient, they are not equally productive. This apparent inconsistency is due to the effects of scale.

Firm A is operating in the increasing returns to scale portion of the production frontier. It could become more productive by increasing its scale of operation towards point B. Point C is operating in the decreasing returns to scale portion of

[26] For a more formal approach and definition of scale efficiency, see Balk (2001).

the production frontier. It could become more productive by decreasing its scale of operation towards point B.

The firm operating at point B is unable to become more productive by changing its scale of operation. It is said to be operating at the *most productive scale size* (MPSS) or equivalently at the *technically optimal productive scale* (TOPS). Visually, this is the point on the production frontier at which a ray from the origin is tangential to the production frontier. This TOPS point can be defined mathematically as

$$\text{TOPS} = \max\{y / x \mid (x, y) \in S\}, \tag{3.14}$$

which is equivalent to finding the (feasible) production point that maximises productivity. The ray that passes through the TOPS point is often called the CRS technology.[27]

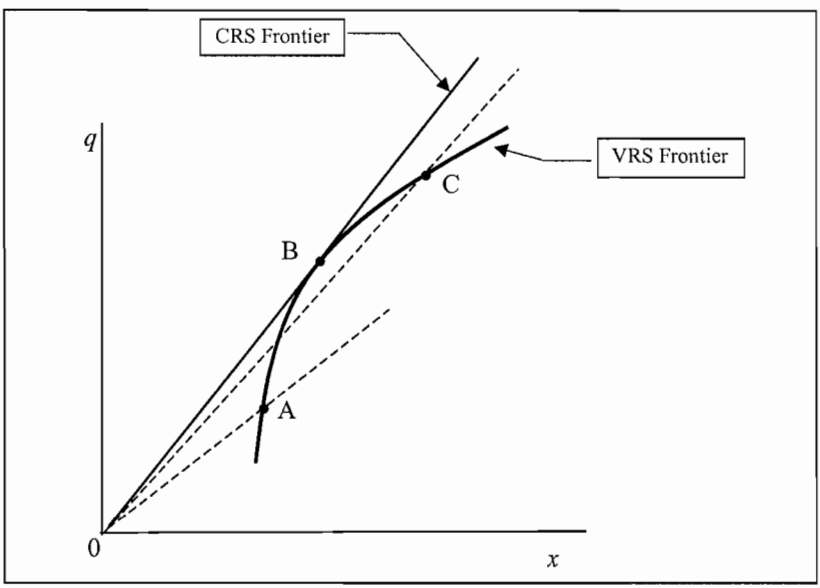

Figure 3.9 The Effect of Scale on Productivity

A *scale efficiency* measure can be used to indicate the amount by which productivity can be increased by moving to the point of TOPS. To illustrate a scale efficiency measure, we reproduce the technology from Figure 3.9 in Figure 3.10, where we depict a technically inefficient firm operating at the point D, and describe how scale efficiency can be calculated using an input orientation.[28] First, from this figure, it is clear that the productivity of firm D (as reflected in the slope of the ray from the origin) could be improved by moving from point D to point E on the VRS

[27] When one considers a production technology involving more inputs and outputs, a TOPS point can be defined for each ray (i.e., each unique mix of inputs and mix of outputs). Thus, the CRS technology is sometimes called a "cone" technology, because its shape is cone-like in the three-dimensional case.
[28] An output-orientated scale efficiency measure can be defined in an analogous manner.

frontier (i.e., removing technical inefficiency), and it could be further improved by moving from the point E to the point B (i.e., removing scale inefficiency).

It is easy to show that the ratio of the slope of the ray 0D to the slope of the ray 0E is equal to the ratio GE/GD, and that the ratio of the slope of the ray 0E to the slope of the ray 0F (which also equals the slope of the ray 0B) is equal to the ratio GF/GE. Thus, we can clearly use distance measures to calculate these productivity differences.

That is, the technical efficiency of firm D relates to the distance from the observed data point to the VRS technology and is equal to the ratio

$$TE_{\text{VRS}} = \text{GE/GD}.$$

Furthermore, the scale efficiency of firm D relates to the distance from the technically efficient data point, E, to the CRS (or cone) technology and is equal to

$$SE = \text{GF/GE}.$$

In the DEA literature, the SE measure is usually not obtained directly, but is calculated indirectly by noting that, if one calculates the distance from the observed data point to the CRS technology (which some authors call a "CRS TE score")

$$TE_{\text{CRS}} = \text{GF/GD},$$

it can then be used to calculate the SE score residually as

$$SE = TE_{\text{CRS}}/TE_{\text{VRS}} = (\text{GF/GD})/(\text{GE/GD}) = \text{GF/GE}.$$

Furthermore, DEA papers often also report the CRS-*TE* measure since it provides a measure of the overall or aggregate productivity improvement that is possible, if the firm is able to alter its scale of operation.[29]

To illustrate these concepts with some rough numbers, consider Figure 3.10 and assume that the data relate to a car wash industry, where each firm is a single person and the input is hours of labour and the output is number of clean cars. Given that firm D washes 10 cars in 10 hours, and that points E and F correspond to 8 hours and 5 hours, respectively, we find that TE_{VRS}=8/10=0.8, SE=5/8=0.625 and, thus, the overall possible productivity improvement is TE_{CRS}=5/10=0.5.

The measurement of scale efficiency in the multi-input, multi-output case is a generalisation of the above concepts. For a particular firm using an input vector, **x**, to produce an output vector, **y**, the concept of TOPS relates to finding a point of maximum productivity on the production frontier, subject to the constraint that the input and output *mixes* cannot be altered, but the *scale* of these vectors can.

Visually, this involves finding all points (δ**x**, λ**y**) on the surface of the production technology, where δ and λ are non-negative scalars. These points produce a two-

[29] Given that a firm is usually unable to alter its scale of operation in the short run, one could view the VRS *TE* score as a reflection of what can be achieved in the short run and the CRS TE score as something that relates more to the long run.

dimensional function similar to that in Figure 3.10. One could then obtain the TOPS point, *corresponding to those particular input and output mixes.*

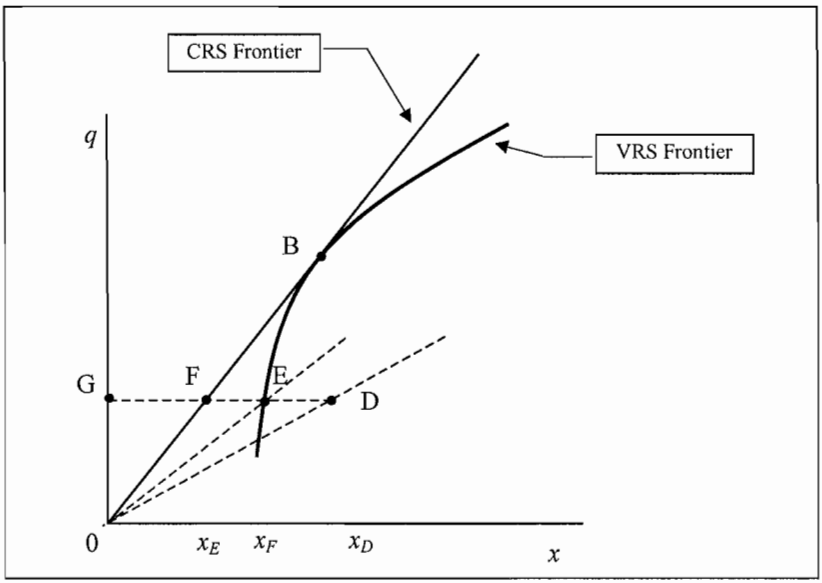

Figure 3.10 Scale Efficiency

This description is a useful visual tool, but using it in practice would be much too time-consuming, because this process needs to be repeated for each firm in the sample. It is much easier to use distance function measures to directly obtain the SE measures.

Following Färe, Grosskopf and Roos (1998), we can define an input-orientated measure of scale efficiency for a firm operating at a given input vector, **x**, and an output vector, **q**, as:

$$SE(\mathbf{x},\mathbf{q}) = \frac{d_i(\mathbf{x},\mathbf{q}|VRS)}{d_i(\mathbf{x},\mathbf{q}|CRS)} = \frac{TE_{CRS}}{TE_{VRS}}. \tag{3.15}$$

The scale efficiency measure in (3.15) can be made operational if the VRS and CRS technologies can be identified so that the distances involved in the definition can be measured appropriately. The calculation of scale efficiency measures using DEA methods is discussed in Chapter 6.

3.5 Measuring Productivity and Productivity Change

The focuses of Chapter 2 and the preceding sections of this chapter have essentially been on developing an appropriate framework and a tool-kit for measuring productivity and productivity change as a part of performance measurement. In this

section, we describe how tolls in this kit can be used in measuring *productivity* and *productivity change*. Productivity is essentially a level concept and measures of productivity can be used in comparing performance of firms at a given point of time. In contrast, productivity change refers to movements in productivity performance of a firm or an industry over time.

3.5.1 Measuring and comparing productivity levels

Measuring productivity is quite simple when only a single output is produced with a single input. In this case, output per unit of input is a comprehensive measure of the level of productivity and it can be used in comparing the performance firms or industries. However, it is a little bit more complex when multiple outputs are produced using multiple inputs. In this case, productivity is often measured using *partial productivity measures* such as output per worker or per hour worked or output per hectare. Though commonly used, partial productivity measures are of limited use and can potentially mislead and misrepresent the performance of a firm. Such measures are often used in growth accounting studies where a researcher wishes to account for differences in labour productivity through changes in capital intensity and other factors. *Multifactor* or *total factor productivity* (MFP or TFP) measures account for the use of a number of factor inputs in production and, therefore, are more suitable for performance measurement and comparisons across firms and for a given firm over time. In the presence of multiple outputs and inputs, total factor productivity may be defined as a ratio of aggregate output produced relative to aggregate input used. Aggregation of outputs and inputs immediately gives rise to index number problems.

A simple TFP measure for firms with multiple outputs and multiple inputs is to look at the profitability of a firm, defined as the revenue of the firm divided by its input cost.[30] Suppose we have two firms producing output vectors $\mathbf{q}_1$ and $\mathbf{q}_2$ using inputs $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively. Suppose the corresponding output and input price vectors are given by $(\mathbf{p}_1, \mathbf{p}_2)$ and $(\mathbf{w}_1, \mathbf{w}_2)$. Then the *profitability ratios* of firms 1 and 2 are given by

$$\pi_1 = \frac{\mathbf{p}_1'\mathbf{q}_1}{\mathbf{w}_1'\mathbf{x}_1} = \frac{\sum_{m=1}^{M} p_{m1}\, q_{m1}}{\sum_{k=1}^{K} w_{k1}\, x_{k1}} \quad \text{and} \quad \pi_2 = \frac{\mathbf{p}_2'\mathbf{q}_2}{\mathbf{w}_2'\mathbf{x}_2} = \frac{\sum_{m=1}^{M} p_{m2}\, q_{m2}}{\sum_{k=1}^{K} w_{k2}\, x_{k2}}. \tag{3.16}$$

A measure of relative performance is given by the ratio, π_2/π_1. Though π_1 and π_2 are scalar measures of total or multifactor productivity, a strict comparison of π_1 and π_2 is difficult since the output and input prices faced by these firms are different.

[30] An alternative measure of performance could simply be profit, revenue minus costs. It is easy to see that profit and profitability are closely inter-related. Use of profitability, since it is in a ratio form, makes it more suitable for purposes of studying its growth.

The only option here is to adjust the value aggregates in equation (3.16) for differences in price levels. Such an adjustment requires that the value aggregates in the numerator and the denominator of equation (3.16) are deflated by suitable price deflators or price index numbers.

In the simple case of firms with a single input and single output, we have the data for the two firms given by (p_1, q_1, w_1, x_1) and (p_2, q_2, w_2, x_2). In this case, a comparison of the profitability ratios is given by:

$$\frac{\pi_2}{\pi_1} = \frac{(p_2 \cdot q_2 / w_2 \cdot x_2)}{(p_1 \cdot q_1 / w_1 \cdot x_1)} = \frac{(p_2 \cdot q_2 / p_1 \cdot q_1)}{(w_2 \cdot x_2 / w_1 \cdot x_1)}. \tag{3.17}$$

Equation (3.17) provides a comparison of profitability between firms 2 and 1. If we make an adjustment for differences in prices faced by firms 1 and 2, by dividing the numerator of equation (3.17) by (p_2/p_1) and the denominator by (w_2/w_1), the profitability ratio in equation (3.17) reduces to:

$$\frac{\pi_2{}^*}{\pi_1{}^*} = \frac{(p_2 \cdot q_2 / p_1 \cdot q_1)/(p_2 / p_1)}{(w_2 \cdot x_2 / w_1 \cdot x_1)/(w_2 / w_1)} = \frac{(q_2 / q_1)}{(x_2 / x_1)} = \frac{(q_2 / x_2)}{(q_1 / x_1)}. \tag{3.18}$$

Equation (3.18) has some interesting features. It shows that a comparison of productivity of firms 1 and 2 using the profitability ratio simply reduces to a ratio of output level differences measured by q_2/q_1 to the input ratio x_2/x_1. If firm 2 produces 50 per cent more output than firm 1 and uses only 25 per cent more input, then a measure of the productivity level of firm 2 relative to 1 is given by 1.50/1.25 = 1.2. This is consistent with what is understood by productivity differences between firms. The last part of equation (3.18) also shows that the relative profitability is indeed a ratio of productivity of firms 2 and 1 as measured by output per unit of input.

It is easy to see that it is not a straightforward exercise to adjust for price level differences when there are multiple inputs and outputs. In such a case, we need to make use of appropriate price deflators constructed using an appropriate index number methodology.[31] The case of multiple inputs and outputs, within the context of measuring productivity change over time, is considered further in the next subsection.

We point out that it is easy to identify the main sources of differences in the profit ratios of two firms. The first and foremost is the differences in prices paid for inputs and outputs by the two firms. So, a given firm can be more profitable in nominal terms if it enjoys favourable output prices relative to input prices, or favourable terms of trade. In assessing productivity performance, we recommend

[31] The theory and practice of constructing input and output price index numbers is considered in detail in Sections 4.6.1 and 4.6.2. At this stage, it is sufficient if the reader is aware of the need for adjustment for price level differences when comparing the performance of firms 1 and 2 using the profitability ratio.

that the price effects are removed. Once the productivity levels are compared using real output and real input measures (derived by deflating nominal aggregates by appropriate price index numbers), then the profitability ratio essentially depends upon the relative efficiency of the two firms. If both firms operate under the same technology, since level comparisons are made at a given point of time, then the profitability-based productivity measure depends upon the *technical, allocative* and *scale efficiency* levels[32] of the two firms in question.

3.5.2 Measuring Productivity Change and the Total Factor Productivity Index

In this subsection, we consider the problem of measuring change in the productivity of a firm or an industry from one period to another. Here, we maintain a subtle distinction between measuring productivity of a firm and that of measuring change in productivity. In the case of firms producing multiple outputs using multiple inputs, we represent change or growth (or decrease) of productivity by a *total factor productivity* (TFP) or a *multifactor productivity index* (MFP). We use TFP and MFP interchangeably, although there is a subtle difference between what each of them may include.[33]

Let us consider the problem of measuring productivity change for a firm from period (or year) s to period t. We assume that the firm makes use of the state of knowledge, as represented by production technologies S^s and S^t in periods s and t. Suppose the firm produces outputs $\mathbf{q}_s$ and $\mathbf{q}_t$ using inputs $\mathbf{x}_s$ and $\mathbf{x}_t$, respectively. In some cases, we may have information on output and input prices, which are represented by output price vectors, $\mathbf{p}_s$ and $\mathbf{p}_t$, and input vectors, $\mathbf{w}_s$ and $\mathbf{w}_t$, in periods s and t, respectively.

Given these data on this firm, how do we measure productivity change? There are several simple and intuitive approaches we can use in deriving meaningful measures of productivity change. We consider four possible alternatives:

- The first approach is to simply use a measure of output growth, net of growth in inputs. If output has doubled over the period s to t, and if this output growth was achieved using only a 60 percent growth in input use, we conclude that the firm has achieved productivity growth. Diewert (1992) has attributed this simple approach to Hicks (1961) and Moorsteen (1961) – thus this approach is known as the Hicks-Moorsteen approach.
- The second approach is to extend the profitability approach and measure productivity change using growth in profitability after making appropriate adjustments for movements in input and output prices over the period s to t.
- The third approach, advocated in Caves, Christensen and Diewert (1982a), thus labelled as the CCD approach, is to measure productivity by comparing

[32] These terms were discussed in Section 3.4.

[33] It is a philosophical question as to whether we can ever take into account all the factors influencing output levels, thus MFP may be a more appropriate term to use.

the observed outputs in period s and t with the maximum level of outputs (keeping the output mix constant) that can be produced using $\mathbf{x}_s$ and $\mathbf{x}_t$, operating under the reference technology. With respect to the reference technology, suppose the firm produced 70 per cent of the maximum feasible output for the given input vector, $\mathbf{x}_s$, in period s and, in period t, it produced 30 per cent above the maximum feasible output for the given input vector, $\mathbf{x}_t$, then a measure of productivity change from period s to t is given by the ratio 1.30/0.70 = 1.857.

- Finally, one may use an entirely different approach in measuring productivity change. Suppose we think and identify various sources of productivity growth: technical change; efficiency change; change in the scale of operations; etc. If we can measure these effects separately, then productivity change can then be measured as the product (or sum total) of all these individual effects. Balk (2001) describes this approach and discusses the resulting measure of productivity change with those recommended in the literature. We label this as the *component-based approach* to productivity change measurement.

We now discuss each of these four approaches and examine their inter-relationships and also indicate material in this book that can be used in implementing each of these approaches. We use the following notation in discussing these methods. We let $\mathrm{TFP}_{s,t}$ denote the total factor productivity index measuring productivity change from period s to t.

An important requirement for the TFP Index

Irrespective of which approach is employed in measuring the TFP index, it is important that it satisfies the following property. If a firm produces the same output quantities in both periods s and t but the input use is *decreased* by a proportion then the TFP index should increase accordingly. If the inputs are reduced by 25 per cent (outputs are produced with only 75 per cent of the original inputs) then the TFP index should be equal to 1/0.75. Similarly, if the outputs are increased by a given percentage, keeping the inputs fixed, then the TFP index should increase by the same percentage. If all the outputs increase by 30 per cent over the period s to t with input use remaining the same then the TFP index should be equal to 1.3.

Suppose we represent the TFP index by a function $\mathrm{F}(\mathbf{x}_t, \mathbf{q}_t, \mathbf{x}_s, \mathbf{q}_s)$ with period s and period t input and output vectors as arguments. Then, we expect any meaningful TFP index to satisfy the property:

$$\mathrm{F}(\lambda\mathbf{x}_s, \mu\mathbf{q}_s, \mathbf{x}_s, \mathbf{q}_s) = \mu/\lambda \ \textit{for all} \ \mu, \lambda > 0 . \tag{3.19}$$

This means that the index is homogeneous of degree +1 in $\mathbf{q}$ and -1 in $\mathbf{x}$. While selecting a formula or an approach to compute the TFP index, we need to ensure that the resulting index satisfies equation (3.19).

Hicks-Moorsteen TFP (HM TFP) Index

The Hicks-Moorsteen index, attributed to Hicks (1961) and Moorsteen (1961) by Diewert (1992),[34] represents a fairly simple TFP index that measures the growth in output, net of growth in inputs. This is the first of the four approaches we listed above. If output growth and input growth are measured using output and input quantity index numbers, then the HM TFP index is given by:

$$\text{HM TFP Index} = \frac{\text{Growth in output}}{\text{Growth in input}} = \frac{\text{Output quantity index}}{\text{Input quantity index}}. \tag{3.20}$$

The HM index can be made operational once appropriate measures of output and input growth are selected. A range of index number formulae are available for this purpose. These are discussed in Chapter 4 – see Section 4.6. This index is also closely related to the index that is based on profitability ratios and the TFP index that is based on the CCD approach.

Though this index is easy to measure and interpret, it is quite difficult to identify the main sources of productivity growth. Suppose, we find that productivity has grown by 10 per cent, do we attribute this to technical change or to improvements in efficiency? The HM index does not have a conceptual framework that underpins a decomposition of TFP growth estimate[35].

TFP Index Based on the Profitability Ratio

Let R_s, R_t, C_s and C_t, respectively, represent the observed revenues and costs of a given firm in periods s and t. The data on input and output quantities and their prices are given by $(\mathbf{x}_s, \mathbf{q}_s, \mathbf{p}_s)$ and $(\mathbf{x}_s, \mathbf{q}_s, \mathbf{w}_s)$ for period s and $(\mathbf{x}_t, \mathbf{q}_t, \mathbf{p}_t)$ and $(\mathbf{x}_t, \mathbf{q}_t, \mathbf{w}_t)$ for period t. The TFP index that is based on the profitability ratio is measured using revenues and costs after adjusting for changes from period t to period s. Let R_s^*, R_t^*, C_s^* and C_t^* represent revenues and costs for the firm in periods s and t, respectively, after adjusting for price changes from period s to period t. Then the TFP index is defined as

$$\text{TFP index} = \frac{R_t^* / R_s^*}{C_t^* / C_s^*} = \frac{\left(R_t / R_s\right) / \text{output price index}}{\left(C_t / C_s\right) / \text{input price index}}. \tag{3.21}$$

where appropriate index formulae are used in measuring price changes from period s to period t.

Since the TFP measure in equation (3.21) does not contain any price effects, the main sources of TFP change over periods s and t can be attributed to technical

[34] This index is also discussed in Bjuerk (1996) and Färe *et al.* (1998).

[35] If the output and input quantity indexes are defined using the Malmquist quantity index (see Section 4.6) then it would be possible to provide a decomposition of the HM TFP index.

change and (technical, allocative and scale) efficiency changes over this period. Issues relating to the actual computation are further considered in Section 4.6.

Malmquist TFP Index

The Malmquist TFP index was first introduced in two very influential papers by Caves, Christensen and Diewert (hereafter, CCD) (1982a, 1982b). In these papers, CCD defined the TFP index using Malmquist input and output distance functions, and thus the resulting index has come to be known as the *Malmquist TFP index*. The method of using these distance functions in defining the TFP index is due to the approach proposed by CCD.

Malmqust TFP index numbers make use of the third approach that is outlined at the beginning of this section. The index is constructed by measuring the radial distance of the observed output and input vectors in periods s and t, relative to a reference technology. As the distances can be either *output orientated* or *input orientated*, the Malmqust TFP indices differ according to the orientation used.[36] These are discussed below.

Output-Orientated TFP Indices

The output-orientated productivity measures focus on the maximum level of outputs that could be produced using a given input vector and a given production technology relative to the observed level of outputs. This is achieved using the output distance functions that are defined in Section 3.2. For purposes of exposition, we concentrate on the period-s Malmquist productivity index, which is given by

$$m_o^{\,s}\left(\mathbf{q}_s,\mathbf{q}_t,\mathbf{x}_s,\mathbf{x}_t\right)=\frac{d_o^{\,s}\left(\mathbf{q}_s,\mathbf{x}_s\right)}{d_o^{\,s}\left(\mathbf{q}_s,\mathbf{x}_s\right)}. \tag{3.22}$$

If we assume that the firm is *technically* efficient in both periods, then $d_o^s(\mathbf{q}_s, \mathbf{x}_s) = 1$, and so[37]

$$m_o^{\,s}\left(\mathbf{q}_s,\mathbf{q}_t,\mathbf{x}_s,\mathbf{x}_t\right)=d_o^{\,s}\left(\mathbf{q}_t,\mathbf{x}_t\right). \tag{3.23}$$

Equation (3.23) shows that $m_o^s(\mathbf{q}_t, \mathbf{q}_s, \mathbf{x}_t, \mathbf{x}_s)$ is the minimal output-deflation factor, such that the deflated-output vector for the firm in period t, $\mathbf{q}_t/[m_o^s(\cdot)]$, and the input vector, $\mathbf{x}_t$, are just on the production surface of the technology in period s. If firm t has a higher level of productivity than is implied by the period-s technology then $m_o^s(\cdot) > 1$.

[36] These two alternative approaches result in the same numerical measure if the technology in periods s and t exhibit the property of global constant returns to scale (CRS).

[37] Equation (3.23) shows that the Malmquist productivity index, in this case, is simply the output distance function defined with respect to period-s technology.

We can similarly define an output-orientated Malmquist productivity index based on period-t technology

$$m_o^{\,t}(\mathbf{q}_s,\mathbf{q}_t,\mathbf{x}_s,\mathbf{x}_t)=\frac{d_o^{\,t}(\mathbf{q}_t,\mathbf{x}_t)}{d_o^{\,t}(\mathbf{q}_s,\mathbf{x}_s)}\,. \tag{3.24}$$

If the firm is *technically efficient in period t* then $d_o^t(\mathbf{q}_t,\mathbf{x}_t) = 1$.

Since the Malmquist productivity index can be defined using period-*s* technology as well as period-*t* technology, the Malmquist TFP index is defined as the geometric average of the two indices based on period-*t* and period-*s* technologies. Thus the output-orientated Malmquist productivity index is given by:

$$m_o\,(\mathbf{q}_s,\mathbf{q}_t,\mathbf{x}_s,\mathbf{x}_t)=\left[m_o^s\,(\mathbf{q}_s,\mathbf{q}_t,\mathbf{x}_s,\mathbf{x}_t)\cdot m_o^t\,(\mathbf{q}_s,\mathbf{q}_t,\mathbf{x}_s,\mathbf{x}_t)\right]^{0.5}. \tag{3.25}$$

We note that the Malmquist TFP index, defined in equation (3.25), requires the computation of four distance functions, namely, $d_o^s(\mathbf{q}_s, \mathbf{x}_s)$, $d_o^t(\mathbf{q}_t,\mathbf{x}_t)$, $d_o^s(\mathbf{q}_t,\mathbf{x}_t)$ and $d_o^t(\mathbf{q}_s, \mathbf{x}_s)$. In order to compute these distance functions, we need to have a description of the production technologies in periods *s* and *t*. If we have very limited data, such as only observed output and input quantities in periods *s* and *t*, then we have to use the index number approach that is discussed in Section 4.6. If we have access to data on a cross-section of firms in periods *s* and *t* then we can use the data envelopment analysis (DEA) approach that is discussed in Chapters 6 and 7 or the stochastic frontier analysis (SFA) that is described in Chapters 8 to 10.

Input-Orientated TFP Indices

The input-orientated productivity focuses on the level of inputs necessary to produce observed output vectors $\mathbf{q}_s$ and $\mathbf{q}_t$ under a reference technology. Suppose we use period-s technology as the reference technology, then the period-*s* input orientated Malmquist productivity index for periods *s* and *t* is defined as:

$$m_i^{\,s}(\mathbf{q}_s,\mathbf{q}_t,\mathbf{x}_s,\mathbf{x}_t)=\frac{d_i^{\,s}(\mathbf{q}_t,\mathbf{x}_{tt})}{d_i^{\,s}(\mathbf{q}_s,\mathbf{x}_s)}. \tag{3.26}$$

If we assume that the firm is *technically efficient*, in both periods, then $d_i^s(\mathbf{q}_s, \mathbf{x}_s) = 1$, and so

$$m_i^{\,s}(\mathbf{q}_s,\mathbf{q}_t,\mathbf{x}_s,\mathbf{x}_t)=d_i^{\,s}(\mathbf{q}_t,\mathbf{x}_t). \tag{3.27}$$

We can similarly define the input-orientated Malmquist productivity index, based on period-t technology, as

$$m_i^t(\mathbf{q}_s, \mathbf{q}_t, \mathbf{x}_s, \mathbf{x}_t) = \frac{d_i^t(\mathbf{q}_t, \mathbf{x}_s)}{d_i^t(\mathbf{q}_s, \mathbf{x}_s)} . \tag{3.28}$$

If the firm is *technically efficient in period t,* then $d_i^t(\mathbf{q}_t, \mathbf{x}_t) = 1$.

Since the Malmquist input-orientated index can be defined using period-s or period-t technology as the reference technology, CCD defined the input-orientated Malmquist TFP index as:

$$m_i(\mathbf{q}_s, \mathbf{q}_t, \mathbf{x}_s, \mathbf{x}_t) = \left[m_i^s(\mathbf{q}_s, \mathbf{q}_t, \mathbf{x}_s, \mathbf{x}_t) \cdot m_i^t(\mathbf{q}_s, \mathbf{q}_t, \mathbf{x}_s, \mathbf{x}_t)\right]^{0.5} . \tag{3.29}$$

If we wish to compute the Malmquist TFP index in equation (3.29), we need to compute the four difference distances that are involved in equations (3.26) and (3.28). If the firm is assumed to be *technically efficient*, then only two distances need to be computed. We encounter problems in the calculations of the Malmquist TFP indices. That is, in order to compute these indices, we need to know the functional form for the distances functions as well as the numerical values of the relevant parameters or, equivalently, a description of the underlying technology. These require firm-level data on inputs and outputs in periods s and t as well as frontier methods that do not require the assumption of technical efficiency of the firms observed.

As the Malmquist TFP index has become a commonly-used measure of productivity change and has gained prominence in the literature, we discuss a few analytical properties of the Malmquist productivity index. However, actual empirical implementation of the Malmquist index is considered in later chapters.

Malmquist TFP and Orientation

We note that the Malmquist TFP index can give different numerical values depending on the type of orientation used. We get different values under ouput- and input-orientated approaches. If the underlying production technology exhibits constants returns to scale (CRS) in both periods, then the input- and output-orientated Malmquist TFP indices coincide.

Malmquist and HM TFP Indices

These two indices are defined using an entirely different conceptual framework, but these indices are interrelated through the following result. The Malmquist TFP indices that are described in equations (3.25) and (3.29), are equal to the Hicks-Moorsteen index, defined in equation (3.20), if and only if the technology is

Moorsteen index, defined in equation (3.20), if and only if the technology is inversely homothetic[38] and exhibits constant returns to scale. Proof of this result is in Färe, Grosskopf and Roos (1996).

Malmquist TFP Index and Technical Inefficiency

When we described the Malmquist TFP index, we observed that it is considerably simplified when the firm is *technically efficient* in both periods s and t. However, if the firm is inefficient then it is possible that observed productivity improvements (change) reflected in the Malmquist TFP index could be the result of improvements in *technical efficiency* (efficiency change) and/or in the underlying *production technology* (technical change). In this case, it is possible to decompose the Malmquist TFP index into two components, one measuring efficiency change and the other measuring technical change. In the following discussion, we simply focus on the output-orientated Malquist TFP index.

The output-orientated Malmquist TFP index in equation (3.25) is the geometric mean of the indices based on period-s and period-t technologies, given in equations (3.22) and (3.24). The index is given by:

$$\begin{aligned} m_o(\mathbf{q}_s,\mathbf{q}_t,\mathbf{x}_s,\mathbf{x}_t) &= \left[m_o^s(\mathbf{q}_s,\mathbf{q}_t,\mathbf{x}_s,\mathbf{x}_t)\times m_o^t(\mathbf{q}_s,\mathbf{q}_t,\mathbf{x}_s,\mathbf{x}_t)\right]^{0.5} \\ &= \left[\frac{d_o^s(\mathbf{x}_t,\mathbf{q}_t)}{d_o^s(\mathbf{x}_s,\mathbf{q}_s)}\times\frac{d_o^t(\mathbf{x}_t,\mathbf{q}_t)}{d_o^t(\mathbf{x}_s,\mathbf{q}_s)}\right]^{0.5}. \end{aligned} \tag{3.30}$$

It is common to observe some degree of inefficiency in the operations of most firms. Hence, assuming that $d_o^s(\mathbf{x}_s,\mathbf{q}_s) \leq 1$ and $d_o^t(\mathbf{x}_t,\mathbf{q}_t) \leq 1$ is likely to be more realistic. Where technical inefficiency is present, the output-orientated Malmquist TFP in equation (3.30), can be rewritten as:

$$m_o(\mathbf{q}_s,\mathbf{q}_t,\mathbf{x}_s,\mathbf{x}_s) = \frac{d_o^t(\mathbf{x}_t,\mathbf{q}_t)}{d_o^s(\mathbf{x}_s,\mathbf{q}_s)}\left[\frac{d_o^s(\mathbf{x}_t,\mathbf{q}_t)}{d_o^t(\mathbf{x}_t,\mathbf{q}_t)}\times\frac{d_o^s(\mathbf{x}_s,\mathbf{q}_s)}{d_o^t(\mathbf{x}_s,\mathbf{q}_s)}\right]^{0.5}, \tag{3.31}$$

where the ratio outside the square brackets measures the change in the output-orientated measure of technical efficiency between periods s and t, and the geometric mean of the two ratios inside the square brackets captures the shift in technology between the two periods, evaluated at $\mathbf{x}_s$ and $\mathbf{x}_t$.

That is, the efficiency change is equivalent to the ratio of the Farrell technical efficiency in period t to the Farrell technical efficiency in period s. The remaining part of the index in equation (3.31) is a measure of technical change. It is the

[38] See Färe and Primont (1995) for the definition of inverse homotheticity.

geometric mean of the shift in technology between the two periods, evaluated at $\mathbf{x}_t$ and also at $\mathbf{x}_s$. Thus the two terms in equation (3.31) are:

$$\text{Efficiency change} = \frac{d_o^t(\mathbf{x}_t, \mathbf{q}_t)}{d_o^s(\mathbf{x}_s, \mathbf{q}_s)} \tag{3.32}$$

and

$$\text{Technical change} = \left[\frac{d_o^s(\mathbf{x}_t, \mathbf{q}_t)}{d_o^t(\mathbf{x}_t, \mathbf{q}_t)} \times \frac{d_o^s(\mathbf{x}_s, \mathbf{q}_s)}{d_o^t(\mathbf{x}_s, \mathbf{q}_s)}\right]^{0.5} \tag{3.33}$$

This decomposition is illustrated in Figure 3.11 where we have depicted a constant returns-to-scale technology involving a single input and a single output. The firm produces at the points D and E in periods s and t, respectively. In each period, the firm is operating below the technology for that period. Hence, there is technical inefficiency in both periods. Using equations (3.30) and (3.31), we obtain:

$$\text{Efficiency change} = \frac{q_t / q_c}{q_s / q_a} \tag{3.32}$$

and

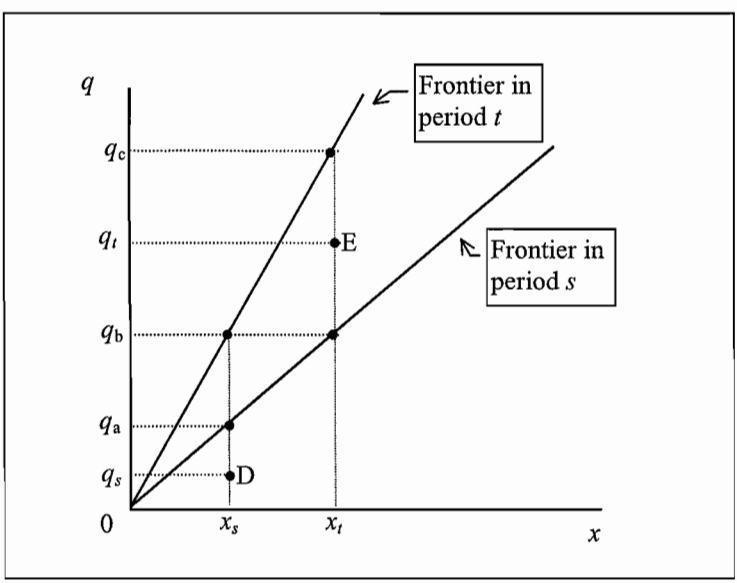

Figure 3.11 Malmquist Productivity Indices

$$\text{Technical change} = \left[\frac{q_t / q_b}{q_t / q_c} \times \frac{q_s / q_a}{q_s / q_b}\right]^{0.5}. \tag{3.33}$$

Given that suitable panel data are available, the various distance functions in equation (3.31) can be directly calculated. Two methods (DEA and SFA) that may be used to calculate these distance functions are discussed in Chapter 11. These two techniques form the subject matter for the next four chapters.

Malmquist TFP Index and Returns to Scale Properties

The Malmquist TFP index, as defined by Caves, Christensen and Diewert (1982a), is not based on specific assumptions about the returns-to-scale properties of the production technologies that underpin the observed output and input quantity vectors. All the distances involved in both input-orientated and output-orientated Malmquist TFP indices can be computed when the technology exhibits variable returns to scale or constant returns to scale. There is no need to specify *a priori* the nature of the production technology.

From the discussion above, it is evident that the Malmquist TFP index captures two important sources of productivity change, gains through *efficiency change* and *technical change*. The question that arises is whether there are *other* sources of productivity growth that are not captured by the Malmquist TFP index. Hence, the returns-to-scale properties of the technology play a role. If the production technology exhibits *constant returns to scale* then there are only two sources of productivity growth: *efficiency change* and *technical change*. For example, Färe *et al.* (1994) use this decomposition in studying productivity performance of OECD countries[39]. However, if the production technology exhibits variable returns to scale there are two other sources of productivity growth.

First, it may be possible that, even in the absence of any technical change (technology remains the same in both periods) and, in the case where the firm under consideration is technically efficient in both periods, there is scope for improving productivity by improving the scale of operations or through improvements in scale efficiency. But the Malmquist TFP index, in this case, shows no productivity growth and the index value is equal to 1. This means that the TFP index fails to capture productivity improvements through improvements in scale efficiency. Grifell-Tatjé and Lovell (1999) propose a generalised Malmquist productivity index that captures productivity gains through improvements in scale efficiency.[40]

---

[39] In fact Färe *et.* (1994) also decompose efficiency change into pure technical efficiency change and scale efficiency change. However, their technical change measure is based on the CRS representations of the technology.

[40] By the definition of scale efficiency in Section 3.4, it is clear that no scale efficiency improvements are possible when the underlying technology is CRS in both periods.

Another source of productivity improvements is through the ability of the firm to exploit possible economies of scope achieved through variations in the output-mix and the input-mix. Balk (2001) demonstrates that productivity change could occur through variations in output-mix, measured using the *output-mix effect (OME)* and the *input-mix effect (IME)*. In the case of a single input and a single output, these effects are equal to 1. Further, if the technology exhibits CRS then this effect is again equal to 1.

In summary, it is clear that, if the technology exhibits CRS, then efficiency change and technical change are the only two sources of productivity change and these are captured by the Malmquist TFP index. However, if a VRS technology is more appropriate then the Malmquist TFP index fails to capture productivity change from all the different sources. However, the standard decomposition of the Malmquist TFP index into technical change and efficiency change components remains valid. This issue is further elaborated below when we discuss the strategy of measuring productivity change by cumulating change through all possible sources of productivity growth.

Malmquist TFP Index and Transitivity

The issue of transitivity has received some attention (see Førsund, 1990; Färe, 1993; Althin, 1995; and Balk and Althin, 1996). Since much of the work on productivity change is based on time-series data, the issue of transitivity has not been a problem. Most studies are content with measuring productivity changes over consecutive years. However, if we are interested in computing productivity indices on cross-sectional elements, such as firms within an industry over a number of years, or even for a given firm over time, then it is necessary to ensure some internal consistency in the results.

Consider the simple scenario where we measure productivity change from period t to period $t+1$ and then also from period $t+1$ to period $t+2$. These productivity indices can be chained to yield a comparison between periods t and $t+2$. Would this index be the same as obtained if we compared period t directly with period $t+2$? The answer is generally in the negative, both from an analytical perspective, as well as from a computational point of view. If we denote the Malmquist TFP index by MTFP, then

$$\text{MTFP}\,(t, t+2) \neq \text{MTFP}\,(t, t+1) \times \text{MTFP}\,(t+1, t+2).$$

The main reason for this possible inconsistency is due to the nature of the underlying production technology. This can be seen from the Malmquist TFP indices in equation (3.25) and (3.29).[41]

[41] We note that both HF TFP index and the TFP index, based on profitability ratios, also do not satisfy the transitivity property.

It is easy to see that in the technical efficiency change measure, the first component is transitive. However, the second component is not transitive unless technical change over time is neutral. Under non-neutral technical change, the output- (or input-) orientated productivity indices depend on the technology under which the various distance functions are derived.

The problem is compounded when panel data sets are used. Balk and Althin (1996) examine the issue of transitivity and propose a transitive index for multilateral comparisons, which, as a by-product, also provides a measure of non-neutrality of the production technology.

The derivation of transitive index numbers is discussed in Chapter 4, where we discuss the Elteto-Koves-Szulc (EKS) method that generates transitive index numbers from non-transitive bilateral comparisons. Hence, it is feasible to generate transitive multilateral Malmquist productivity indices from bilateral indices using the EKS procedure. Even the procedure of Balk and Althin (1996) is somewhat mechanical in its approach. Førsund (2004) provides a summary of the work to date and provides a clear statement of all the issues surrounding the construction of transitive productivity indices. However, Førsund (2004) does not offer any concrete solutions to the problem. Further work is needed for generating a theoretically-meaningful transitive multilateral Malmquist productivity index.

TFP Index – Measurement by Sources of Productivity Change

The three alternative approaches to the measurement of productivity change that are discussed above make use of an intuitively-appealing conceptualisation of productivity change and then offer a measure that can be empirically implemented. Once a numerical measure of productivity change is obtained, an interpretation of the measure is required. Does the measure represent pure technical change or does it also capture efficiency change? This type of approach to productivity measurement may be considered as a *top-down approach*, under which it is possible that some sources of productivity change may not be adequately accounted for and there could be some difficulty in the interpretation of the results.

In this sub-section, we describe an alternative approach to productivity measurement that tries to identify all the sources of productivity change and then constructs a measure of the growth in total factor productivity. This is an approach advocated by Balk (2001). This approach may be considered as a *bottoms-up approach*, which starts with a list of all possible sources of productivity growth and then examines the best possible way of measuring each of these sources and combines them to derive a measure of productivity change. The resulting measure of productivity change should satisfy the basic property, stated in equation (3.19), that the productivity change measure must be homogeneous of degree +1 in outputs and homogeneous of degree -1 in inputs.

Balk (2001) identifies four sources of productivity growth. All these sources are easy to understand and can be seen to be important drivers of productivity change. The first, and the most commonly-considered, source of productivity growth is *technical change* (TC), which results from a shift in the production technology. The second source of productivity growth of a firm could be due to improved efficiency in the firm's ability to use the available technology, which is *efficiency change (EC)*. It is conceivable that a firm could increase its productivity even when there is no technical change by making a more efficient use of its inputs and by operating closer to the technology frontier. The third source is due to improvements in scale efficiency, measured through *scale efficiency change* (SEC). This source refers to improvements in the scale of operations of the firm and its move towards *technologically optimum scale* (TOPS) of operations. These three sources have already been discussed in this chapter and, in the simple case of a single output produced using a single input, these are sufficient to capture all the sources of productivity change from a primal approach to productivity change – an approach based on the primal representation of technology. However, in the case of multi-output and multi-input firms another factor can also result in productivity change, the *output mix effect* (OME) or the *input mix effect* (IME), which measure the effects of changes in the composition of the output and input vectors over periods s and t. This is a source of productivity that has not been commonly discussed in the literature and also does not feature in any of the interpretations or decompositions of the TFP index. Balk (2001) presents a detailed discussion of OME and IME, which forms the basis for the following description of this source-based approach to TFP growth measurement.

In addition to the notation already used, we introduce the concept of *cone technology* associated with a given technology. If S represents the production technology then the cone technology associated with S, denoted by S^*, is defined as the smallest cone constructed out of all the elements of S. Thus, if

$$\mathrm{S} = \left[\text{all } (\mathbf{x},\mathbf{q}) \text{ such that } \mathbf{x} \text{ produces } \mathbf{q}\right]$$

then the cone technology associated with S is defined as:

$$\mathrm{S}^* = \left\{(\lambda\mathbf{x},\lambda\mathbf{q}) \,\middle|\, (\mathbf{x},\mathbf{q}) \in S,\ \lambda > 0\right\}. \tag{3.34}$$

It is easy to see that S* is the set consisting of the rays passing through all the feasible input-output combinations ($\mathbf{x}$, $\mathbf{q}$). If the technology S exhibits constant returns to scale, then S* = S.[42] We denote the distances measured relative to a *cone technology* by $d^*(.)$ and distances with respect to the observed technology by $d(.)$.

[42] We wish to emphasise that S* is essentially an artificial construct derived from a given technology. This means that certain points in S* may not be feasible according to the observed technology S. It is quite straightforward to construct the cone technology when S is known.

Now we focus on each of the four sources of TFP change listed above. We note that three out of four of these have already been dealt with in Section 3.4. Also, all these sources can be considered from an output orientation or from an input orientation. In the discussion below, we focus on output-orientated measures of TFP but input-orientated measures can be similarly obtained. We use the subscript "o" with all the measures to indicate that the output orientation is used.

Technical Change

We measure technical change experienced by a firm through its ability to produce more (or less) with a given vector of input quantities in period t in comparison to the levels feasible in period s. Technical change can be measured relative to a given input and output vector. We consider a pair of vectors ($\mathbf{x}$, $\mathbf{q}$) such that there is at least one non-zero output vector associated with the input vector, $\mathbf{x}$, under the technologies of periods s and t. Then, making use of output distance functions, we can measure technical change by comparing the radial projection of the output vector, $\mathbf{q}$, onto the frontiers of S_s and S_t. TC denotes the technical change measure that is defined as:

$$TC_o^{s,t} = \frac{d_o^t(\mathbf{x},\mathbf{q})}{d_o^s(\mathbf{x},\mathbf{q})}. \tag{3.35}$$

A numerical value for the TC measure of greater than 1 implies that there is technical progress.

We note that the technical change measure is a function of the choice of $\mathbf{x}$ and $\mathbf{q}$. In practice, the most obvious choices are the observed input and output vectors in periods s and t. These two choices result in two measures of technical change and it is difficult to advocate the use of one against the other. So, an average of these two measures is usually taken. Given that productivity change is measured in a ratio form, it is natural to make use of the geometric mean of the two measures. Thus, the recommended measure of technical change is:

$$TC_o^{s,t}(\mathbf{x}_s,\mathbf{q}_s,\mathbf{x}_t,\mathbf{q}_t) = \left[\frac{d_o^t(\mathbf{x}_s,\mathbf{q}_s)}{d_o^s(\mathbf{x}_s,\mathbf{q}_s)} \times \frac{d_o^t(\mathbf{x}_t,\mathbf{q}_t)}{d_o^s(\mathbf{x}_t,\mathbf{q}_t)}\right]^{0.5}. \tag{3.36}$$

The technical change measure in equation (3.36) is the same as the technical change measure associated with the Malmquist TFP index. Basically, we observe technical change in the direction of $\mathbf{q}_s$ and $\mathbf{q}_t$ and a geometric mean of the two indices is the measure in equation (3.36).

We emphasise that all the distance functions are measured relative to the actual technologies in periods s and t.

Technical Efficiency Change

We recall that technical efficiency of an observed pair of inputs and outputs, from an output orientation, is measured by the extent to which the observed output vector could be radially expanded to be on the frontier of the production possibility set associated with the input vector. Thus $d_o^s(\mathbf{x}_s, \mathbf{q}_s)$ and $d_o^t(\mathbf{x}_t, \mathbf{q}_t)$ are measures of technical efficiency in periods s and t, respectively. Then technical efficiency change, *TEC*, is measured by:

$$TEC_o^{s,t}(\mathbf{x}_t, \mathbf{q}_t, \mathbf{x}_t, \mathbf{q}_t) = \frac{d_o^t(\mathbf{x}_t, \mathbf{q}_t)}{d_o^s(\mathbf{x}_s, \mathbf{q}_s)}. \tag{3.37}$$

This measure of technical efficiency change measure is the same as the measure that forms a component of the Malmquist TFP index. We observe, once again, that the distances involved are computed with respect to the observed production technologies in periods s and t.

Scale Efficiency Change

We introduce the concept of scale efficiency in Section 3.4.3. The essential idea is that a firm could increase its productivity by changing the scale of its operations such that the firm operates at a technologically optimal scale (TOPS) of production. So the scale efficiency of a given firm is then measured using the output distance of the observed input-output vectors relative to the *variable returns-to-scale* (VRS) frontier and from the *cone technology* or the *constant returns-to-scale* (CRS) technology that is generated from the observed VRS technology. Thus, the output-orientated scale efficiency measure in period t is defined as:

$$SE_o^t(\mathbf{x}, \mathbf{q}) = \frac{TE_t^*(\mathbf{x}, \mathbf{q})}{TE_t(\mathbf{x}, \mathbf{q})} = \frac{d^{*t}{}_o(\mathbf{x}, \mathbf{q})}{d_o^t(\mathbf{x}, \mathbf{q})}. \tag{3.39}$$

where TE^* denotes technical efficiency measured with respect to the cone technology. This measure of scale efficiency is based on period-t technology for a given input-output combination, $(\mathbf{x}, \mathbf{q})$. The numerical value of scale efficiency is always in the range 0 to 1 and, if it is equal 1, then the firm is scale efficient.

Using the scale efficiency measure in equation (3.39), we can define a measure of output-orientated scale efficiency change as the ratio of the scale efficiency measures associated with input vectors $\mathbf{x}_s$ and $\mathbf{x}_t$, measured with respect to a specific output vector, $\mathbf{q}$. Thus, we can define output-orientated *scale efficiency change* (*SEC*) as:

$$SEC_o^t(\mathbf{x}_s, \mathbf{x}_t, \mathbf{q}) = \frac{SE_o^t(\mathbf{x}_t, \mathbf{q})}{SE_o^t(\mathbf{x}_s, \mathbf{q})}. \tag{3.40}$$

If this ratio is greater than unity, then we must conclude that the input vector $\mathbf{x}_t$ lies closer to the point of technically optimal scale than the input vector $\mathbf{x}_s$ in period s. We also note that the measure satisfies the property of linear homogeneity in the output vector $\mathbf{q}$. Further, it is clear from the definition that the measure is based on period-t technology only and, thus, it is independent of any technical change effects. Further, if period-t technology exhibits global *constant returns to scale* then the scale efficiency change in equation (3.40) is identically equal to 1.

The SEC measure in equation (3.40) depends upon the choice of an output vector and also the reference technology for computing the relevant distance functions in equation (3.39). A natural choice would be to choose period-s and period-t technologies as reference technologies and output vectors $\mathbf{q}_s$ and $\mathbf{q}_t$, respectively. This choice leads to two different numerical measures of scale efficiency change, and we can take a geometric average as an appropriate measure. Thus, we have the following measure of SEC:

$$SEC_o^{s,t}(\mathbf{x}_s, \mathbf{x}_t, \mathbf{q}_s, \mathbf{q}_t) = \left[SEC_o^s(\mathbf{x}_s, \mathbf{x}_t, \mathbf{q}_s) \times SEC_o^t(\mathbf{x}_s, \mathbf{x}_t, \mathbf{q}_t)\right]^{0.5}. \tag{3.41}$$

This is a measure of scale efficiency change that is consistent with our measure of scale efficiency and it is also in line with our measures of technical change and technical efficiency change.

Output Mix Effect

The *output mix effect* (*OME*) is a novel concept that is designed to capture the effect of the composition of the output vector (or output mix) on scale efficiency. The scale efficiency change in equation (3.41) is defined at the two observed output quantities $\mathbf{q}_s$ and $\mathbf{q}_t$ but it does not capture the influence of change in the output mix implicit in these two vectors. Balk (2001) defines a general measure of *OME*, defined using a selected technology, say period-t technology, and a specific input vector $\mathbf{x}$, which is given by:

$$OME^t(\mathbf{x}, \mathbf{q}_s, \mathbf{q}_t) \equiv \frac{SE_o^t(\mathbf{x}, \mathbf{q}_t)}{SE_o^t(\mathbf{x}, \mathbf{q}_s)}. \tag{3.42}$$

The scale efficiencies in the numerator and denominator are measured in the direction of the output vectors $\mathbf{q}_t$ and $\mathbf{q}_s$, respectively. This ratio captures the effect of change in the direction from $\mathbf{q}_s$ to $\mathbf{q}_t$ on scale efficiency – thus providing a measure of the effect of output mix on scale efficiency.

The output mix effect measure in equation (3.42) depends upon the choice of technology (period s or period t) and the choice of input vector $\mathbf{x}$. Given the choice between period-s and period-t technologies and input vectors, we define *OME* as the geometric of the two resulting measures, which is given by:

$$OME^{s,t}(\mathbf{x}_s, \mathbf{x}_t, \mathbf{q}_s, \mathbf{q}_t) = \left[OME^s(\mathbf{x}_s, \mathbf{q}_s, \mathbf{q}_t) \times OME^t(\mathbf{x}_t, \mathbf{q}_s, \mathbf{q}_t)\right]^{0.5}$$
$$= \left[\frac{SE_o^s(\mathbf{x}_s, \mathbf{q}_t)}{SE_o^s(\mathbf{x}_s, \mathbf{q}_s)} \times \frac{SE_o^t(\mathbf{x}_t, \mathbf{q}_t)}{SE_o^t(\mathbf{x}_t, \mathbf{q}_s)}\right]^{0.5} \quad . \quad (3.43)$$

The output mix effect is the last of the sources of productivity change. While all the three sources described are intuitively clear, the output mix effect is a bit complex. It is designed to capture the effect of output composition on scale efficiency change over the periods under consideration.

Using the properties of various scale efficiency measures used in defining OME in equation (3.43), it is possible to show that the output mix effect is equal to 1 in the single-output case. Further, it is also easy to see that if the output mix remains the same over the periods s and t, then *OME* is again equal to 1.

Measure TFP from Various Sources of Productivity Change

Once we have satisfactorily measured productivity change from the various sources, namely, technical change, technical efficiency change, scale efficiency change and, the effect of output mix, we can bring all these components together to measure TFP change over period s to period t, based on the observed data ($\mathbf{x}_s$, $\mathbf{x}_t$, $\mathbf{q}_s$, $\mathbf{q}_t$), we have:

TFP Change = technical change × technical efficiency change
× scale efficiency change × output mix effect

where all component are measured using formulae described above. Thus, we have:

$$TFPC^{s,t}(\mathbf{x}_s, \mathbf{x}_t, \mathbf{q}_s, \mathbf{q}_t) = \left[\frac{d_o^t(\mathbf{x}_s, \mathbf{q}_s)}{d_o^s(\mathbf{x}_s, \mathbf{q}_s)} \times \frac{d_o^t(\mathbf{x}_t, \mathbf{q}_t)}{d_o^s(\mathbf{x}_t, \mathbf{q}_t)}\right]^{0.5} \times \left[\frac{d_o^t(\mathbf{x}_t, \mathbf{q}_t)}{d_o^s(\mathbf{x}_s, \mathbf{q}_s)}\right]$$
$$\times \left[SEC_o^s(\mathbf{x}_s, \mathbf{x}_t, \mathbf{q}_s) \cdot SEC_o^t(\mathbf{x}_s, \mathbf{x}_t, \mathbf{q}_t)\right]^{0.5} \quad (3.44)$$
$$\times \left[OME^s(\mathbf{x}_s, \mathbf{q}_s, \mathbf{q}_t) \cdot OME^t(\mathbf{x}_t, \mathbf{q}_s, \mathbf{q}_t)\right]^{0.5}$$

where we need to substitute for the scale efficiency and output mix effect measures in terms of the all the appropriate distances. After substitutions and algebraic manipulations, we derive the following expression for TFP change in equation (3.44). Following Balk (2001), we have the following expression:

$$TFPC^{s,t}\left(\mathbf{x}_s,\mathbf{x}_t,\mathbf{q}_s,\mathbf{q}_t\right)=\left[\frac{d^{*s}_o\left(\mathbf{x}_t,\mathbf{q}_t\right)}{d^{*s}_o\left(\mathbf{x}_s,\mathbf{q}_s\right)}\times\frac{d^{*t}_o\left(\mathbf{x}_t,\mathbf{q}_t\right)}{d^{*t}_o\left(\mathbf{x}_s,\mathbf{q}_s\right)}\right]^{0.5}. \tag{3.45}$$

where * indicates that the associated distances are measured from the *cone technology* that is generated from the underlying observed technology.

The result in equation (3.45) is quite interesting. We first note that this is very similar to the Malmquist TFP index that is defined in equation (3.19), where the index is defined with respect to the actual technologies in periods s and t. In contrast, the measure derived in this section is defined with respect to the cone technologies associated with periods s and t.

We note the following important properties of the TFP change measure constructed from its sources:

- It is clear that TFP change can be driven by four different factors. Missing any of these factors may misrepresent the TFP change that is evident in the observed data.
- The Malmquist TFP measure is somewhat incomplete since it takes into account TFP change only due to technical change and efficiency change.
- The generalised Malmqist TFP index that is proposed in Grifell-Tatjé and Lovell (1999) ignores the output-mix effect.
- The overall measure of TFP change in Färe *et al.* (1994) coincides with the source-based TFP change measure in equation (3.45), but the decomposition derived by Färe *et al.* (1994) does not coincide with that in equation (3.45) as they measure technical change, not from the observed (possibly, VRS) technology, but from the cone technology that encompasses the actual technology.
- Balk (2001) also discusses other decompositions of productivity change in the literature due to Ray and Desli (1997), Wheelock and Wilson (1999) and Zofio and Lovell (1999) and shows how they relate to the source-based TFP change measure given in equation (3.45).

We conclude this section with the comment that all the sources of productivity change that are described above can also be measured using an input-orientated approach. Usually numerical measures of these components differ under the input- and output-orientated approaches unless the underlying production technology exhibits global CRS. Though the numerical values of the components, and, therefore, their contribution to overall productivity change may differ, the overall measure of TFP change measure remains the same whether an input or output orientation is used,[43] a very useful property indeed! From a practitioner's viewpoint, this result implies that if the interest is on the overall productivity change then the

[43] This result follows from the fact that the source-based TFP change in equation (3.45) uses output-orientated distances and that, when the technology is global CRS, the input-orientated distance function is simply the reciprocal of the output-orientated distance function.

source-based approach gives a unique answer. If interest is in identifying the contribution of each source to overall productivity change, then it is necessary to select the orientation to be used. Selection of the orientation may depend upon whether the firm's manager has the ability to radially expand the outputs or inputs. We return to the problem of selecting orientation in Chapter 6.

3.6 Conclusions

In this chapter, we provide an overview of the conceptual framework that underpins efficiency and productivity measurement. The main objective is to offer a user-friendly explanation of all the concepts involved and, therefore, the rigour of the exposition is necessarily diluted. For a more rigorous mathematical treatment of the material covered in this chapter, the reader is encouraged to read Färe, Grosskopf and Lovell (1994); Färe and Primont (1995); Balk (1998); and Färe, Grosskopf and Russell (1998). The reader should also attempt to read the two very influential papers by Caves, Christensen and Diewert (1982a, 1982b) that paved the way for the Malmquist TFP index.

In summary, we have covered two major groups of concepts that are central to efficiency and productivity measurement. First, we describe a set-theoretic approach to the production technology, in contrast to the production function and the transformation function approach used in Chapter 2 to study production technologies. The set-theoretic description is useful in defining and establishing the properties of input and output distance functions. In the second part of this chapter, we describe a number of efficiency measures, namely, technical, allocative, economic and scale efficiency measures. Using this framework, we proceed to the problem of measuring productivity level differences across firms and productivity change over time. A number of *total factor productivity* (TFP) measures based on: the Hicks-Moorsteen approach, profitability ratios, the Malmquist approach to productivity measurement, and, finally, the component-based approach are described and their inter-relationships highlighted.

Practitioners may be somewhat confused and may wonder as to which of these approaches should be used. The following are a few points to consider during this decision-making process.

- The type of approach to select should depend upon the purpose of measuring productivity levels and change. For example, if a summary measure of productivity changes is required without any need to identify their sources then we suggest the use of the Hicks-Moorsteen approach or a Mamlquist productivity index based on the *cone technology* that envelops the observed production technology. If a more business-oriented approach is appropriate, the use of changes in *profitability ratios* to measure productivity change is advocated.

- The actual measures of productivity levels and TFP growth that are selected should be empirically feasible, i.e., the right kind of data must be available to implement the measure selected. Suppose, we opt to measure TFP change using the Malmquist index associated with the *cone technology*. In order to obtain this measure, a panel data set must be available on a large number of firms over the period under consideration. This means firm-level data on input and output quantities are required for a large number of firms, large enough to provide a good description of the underlying technology. But if data are only available for a single firm over time, then the approach based on the *cone technology* is not feasible. In this case, only the Hicks-Moorsteen approach is feasible. Similarly, if reliable price data are not available, it is not possible to arrive at measures of allocative efficiency or economic efficiency. In the absence of price data, efficiency and productivity measurement is necessarily restricted to the measurement technical efficiency, scale efficiency and technical change.
- There are instances where the issues of scale are not relevant. In such cases, it may be possible to assume that the production technology exhibits constant returns to scale. This is the case when international comparisons of productivity are studied. Since the endowments of countries, in terms of the size of land, population and natural resources, are given, and their size is not a decision variable, it is appropriate to work with the assumption of CRS technology. In this case, the Malmquist TFP index is sufficient because it coincides with the index resulting from the source-based measure of TFP growth.

In conclusion, we express our opinions in the form of advice to the practitioner. If CRS technology is a tenable assumption in a particular instance, we suggest that the Malmquist TFP index be used if panel data are available. If only limited data are available, the Hicks-Moorsteen approach or the *index number approach*, is, in general, the best option. Under the CRS assumption, the index number approach provides measures of TFP change that are close to those of the Malquist TFP index. If the technology exhibits *variable returns to scale*, then we recommend the component-based approach to productivity measurement. This approach, outlined in Balk (2001), is intuitive and is empirically feasible. It is a relatively new approach but it has the appeal of being internally consistent and just as easy as the Malmquist approach. It also eliminates the need to derive a decomposition of the overall TFP change as the components are first identified and computed. Identifying and estimating production frontiers are difficult tasks that require the use of sophisticated techniques. A major portion of this book is devoted to these techniques. Chapters 6 and 7 are devoted to *data envelopment analysis* (DEA). Estimation of stochastic production frontiers and distance functions is the focus for Chapters 8, 9 and 10. The link between these methods and efficiency and productivity measurement is elaborated in Chapter 11. Before we move on to the next chapter, we wish to point out that the main focus of this chapter was on the measurement of firm level efficiency. We have not examined the issue of aggregating firm level efficiencies into a measure of efficiency of the industry or a

production sector. This is an emerging topic which technically challenging. We recommend that interested readers read the newly published work of Färe and Grosskopf (2004) where these aggregation issues are discussed in detail.

4. INDEX NUMBERS AND PRODUCTIVITY MEASUREMENT

4.1 Introduction

Index numbers are the most commonly-used instruments to measure changes in levels of various economic variables. Index numbers relating to various economic phenomena are regularly compiled and published. The consumer price index (CPI), which measures the changes in prices of a range of consumer goods and services, is the most widely-used economic indicator. Other important index numbers include the price deflators for national income aggregates; indices of import and export prices; financial indices such as the All Ordinaries Index (Australia) and the Dow Jones Index (U.S.A.).

The principal aim of this chapter is to provide a simple exposition to various index numbers that are relevant in the context of measuring productivity changes over time and space. Based on the discussion in the preceding chapters, it is evident that measuring productivity changes necessarily involves measuring changes in the levels of output and the associated changes in the input usage. Such changes are easy to measure in the case of a single input and a single output, but are more difficult when multi-input and multi-output cases are considered.

We see three principal areas in productivity measurement where index numbers play a major role. The first and foremost is the use of index numbers in the measurement of changes in total factor productivity (TFP) leading to the popular TFP index numbers. TFP index numbers in turn require separate input and output quantity index numbers.

The second use of index numbers in productivity measurement is an indirect role. It concerns the use of index numbers in generating the required data that can be used in the application of data envelopment analysis (DEA) or in the estimation of the stochastic frontiers. These two techniques are discussed in Chapters 6 to 9. Application of these techniques using very detailed data on inputs and outputs may pose estimation problems from the loss of degrees of freedom[GEB1] due to inclusion of a large number of input and output categories. In most practical applications of these techniques, it is necessary to "aggregate" data into a smaller number of input and output variables. For example, different types of labour inputs are usually aggregated into one single labour input. Outputs of commodities that belong to a particular group are usually aggregated into a single output aggregate for the group. For example, in agriculture, items such as wheat, rice, etc., are grouped to form the output of "cereals". This type of aggregation, which is essentially an intermediate step in the process of assessing efficiency and productivity change, requires the use of index numbers. Usually, such aggregates take the form of "*value aggregates at constant prices*" or just "*constant price series*".[1]

The third area concerns the type of index numbers that are required in handling panel data sets, with price and quantity data over time and space. Comparison of spatial observations usually requires some basic consistency requirements such as "transitivity" and "base invariance". These requirements, in turn, stipulate that the formulae used for the purpose of generating index numbers, of the type discussed in the preceding two paragraphs, need to satisfy some additional requirements.

The principal aim of the first part of this chapter is to familiarise the reader with the various index numbers, such as the Laspeyres, Paasche, Fisher and Törnqvist index numbers, and then to focus on the construction of price and quantity index numbers. Quantity index number formulae are applied to input and output data that lead to quantity index numbers that are, in turn, used in defining the TFP index. The second part of this chapter deals with the micro-economic theoretic foundations for the measurement of input and output index numbers and demonstrate how TFP change, defined in Chapter 3, can be measured using the index numbers discussed in this chapter.

4.2 Conceptual Framework and Notation

An index number is defined as a real number that measures changes in a set of related variables. Conceptually, index numbers may be used for comparisons over time or space or both. Index numbers are used to measure price and quantity changes over time, as well as to measure differences in the levels across firms, industries, regions or countries. Price index numbers may refer to consumer prices, input and output prices, export and import prices, etc., whereas quantity index numbers may be measuring changes in quantities of outputs produced or inputs used by a firm or industry over time or across firms.

[1] Some of these issues are further discussed in Chapter 5 on data and measurement issues.

Index numbers have a long and distinguished history in economics, with some of the most important contributions due to Laspeyres and Paasche, dating back to the late nineteenth century. The Laspeyres and Paasche formulae are still commonly used by national statistical offices around the world. But it is the work of Irving Fisher and his book, T*he Making of Index Numbers,* published in 1922, that recognised the possibility of using many statistical formulae to derive appropriate index numbers. The Tornquist (1936) index is a formula that plays a major role in productivity measurement. Chapter 2 of Diewert and Nakamura (1993) provides an excellent exposition of the historical background to index number construction.

Notation

We use the following notation throughout this chapter. Let p_{mj} and q_{mj} represent the price and quantity, respectively, of the m-th commodity in the M commodities being considered ($m = 1,2,...,M$) in the j-th period ($j = s, t$). Without loss of generality, s and t may refer to two firms instead of time periods, and quantities may refer to either input or output quantities.

Conceptually, all index numbers measure changes in the levels of a set of variables from a reference period. The reference period is denoted at the "base period". The period for which the index is calculated is referred to as the "current period". Let I_{st} represent a general index number for current period, t, with s as the base period. Similarly, let V_{st}, P_{st} and Q_{st} represent value, price and quantity index numbers, respectively.

The General Index Number Problem

The value change from period s to t is the ratio of the value of commodities in periods s and t, valued at respective prices. Thus

$$V_{st} = \frac{\sum_{m=1}^{M} p_{mt} q_{mt}}{\sum_{m=1}^{M} p_{ms} q_{ms}} . \tag{4.1}$$

The index, V_{st}, measures the change in the value of the basket of quantities of M commodities from period s to period t. Obviously, V_{st} is the result of changes in the two components, prices and quantities. While V_{st} is easy to measure, it is more difficult to disentangle the effects of price and quantity changes. We may want to disentangle these effects so that, for example, the quantity component can be used in measuring the quantity change.

If we are operating in a single-commodity world, then such a decomposition is very simple to achieve. We have

$$V_{st} = \frac{p_t q_t}{p_s q_s} = \frac{p_t}{p_s} \times \frac{q_t}{q_s} \ .$$

Here the ratios p_t/p_s and q_t/q_s measure the relative price and quantity changes and there is no index number problem.

In general, when we have $M \geq 2$ commodities, we have a problem of aggregation. The price relative, p_{mt}/p_{ms}, measures the change in the price level of the m-th commodity, and the quantity relative, q_{mt}/q_{ms}, measures the change in the quantity level of the m-th commodity ($m = 1,2,...,M$),.

Now the problem is one of combining the M different measures of price (quantity) changes, into a single real number, called a *price (quantity) index*. This problem is somewhat similar to the problem of selecting a suitable measure of central tendency. In the next two sections, we briefly examine some of the more commonly-used formulae for measuring price and quantity index changes.

4.3 Formulae for Price Index Numbers

We first focus on the price index numbers and then illustrate how these formulae can also be used in the construction of quantity index numbers.

Laspeyres and Paasche Index Numbers

These index numbers represent the most widely used indices in practice. The Laspeyres price index uses the base-period quantities as weights, whereas, the Paasche index uses the current-period weights to define the index.

$$\text{Laspeyres index} = P_{st}^{L} = \frac{\sum_{m=1}^{M} p_{mt} q_{ms}}{\sum_{m=1}^{M} p_{ms} q_{ms}} = \sum_{m=1}^{M} \frac{p_{mt}}{p_{ms}} \times \omega_{ms} \ , \tag{4.2}$$

where $\omega_{ms} = p_{ms} q_{ms} \Big/ \sum_{m=1}^{M} p_{ms} q_{ms}$ is the value share of m-th commodity in the base period.

Equation (4.2) suggests two alternative interpretations. First, the Laspeyres index is the ratio of two value aggregates resulting from the valuation of the base-period quantities at current- and base-period prices. The second interpretation is that the index is a value-share weighted average of the M price relatives. The value shares reflect the relative importance of each commodity in the set involved. The value shares used here refer to the base period.

A natural alternative to the use of base-period quantities, in the definition of the Laspeyres index, is to use current-period quantities. The Paasche index number, which makes use of the current-period quantities, is given by

$$\text{Paasche index} = P_{st}^{P} = \frac{\sum_{m=1}^{M} p_{mt} q_{mt}}{\sum_{m=1}^{M} p_{ms} q_{mt}} = \frac{1}{\sum_{m=1}^{M} \frac{p_{ms}}{p_{mt}} \times \omega_{mt}} . \tag{4.3}$$

The first part of equation (4.3) shows that the Paasche index is the ratio of the two value aggregates resulting from the valuation of period-t quantities at the prices prevailing in periods t and s. Alternatively, the last part of the equation suggests that the Paasche index is a weighted harmonic mean of price relatives, with current-period value shares as weights.

From equations (4.2) and (4.3), it can be seen that the Laspeyres and Paasche formulae, in a sense, represent two extremes, one formula placing emphasis on base-period quantities and the other on current-period quantities. These two indices coincide if the price relatives do not exhibit any variation, that is, if $p_{mt}/p_{ms} = c$, then the Laspeyres and Paasche indices are both equal to the constant c. These indices tend to diverge when price relatives exhibit a large variation. The extent of divergence also depends on quantity relatives and the statistical correlation between price and quantity relatives. Bortkiewicz (1924) provides a decomposition of the Laspeyres-Paasche gap.

The Laspeyres and Paasche indices are quite popular since they are easy to compute and they provide "bounds" for the *true index* that is defined using economic theory (see Section 4.6). Most national statistical agencies use these formulae or some minor variations in computing various indices, such as the CPI. In particular, use of the Laspeyres index is more prevalent. We note that if published price indices are being used for purposes of deflating a given value series, then the resulting deflated series will exhibit definite characteristics depending on which formula is used in constructing the price index numbers. This issue is further elaborated in Section 4.4, where the indirect measurement of quantity changes are discussed.

The Fisher Index

The gap between the Laspeyres and Paasche indices led Fisher (1922) to define a geometric mean of the two indices as a possible index number formula:

$$\text{Fisher index} = P_{st}^{F} = \sqrt{P_{st}^{L} \times P_{st}^{P}} . \tag{4.4}$$

Though the Fisher index is an artificial construct, which has value between the two extremes, it possesses a number of desirable statistical and economic theoretic properties. Diewert (1992) demonstrates the versatility of the Fisher index. In view

of the many favourable properties it has, the Fisher index is also known as the *Fisher ideal index.*

The Törnqvist Index

The Törnqvist index has been used in many total factor productivity studies that have been conducted in the last decade. In this section, we define the Törnqvist price index, while we define the Törnqvist quantity index in the next section. The Törnqvist price index is a weighted geometric average of the price relatives, with weights given by the simple average of the value shares in periods s and t:

$$P_{st}^{T} = \prod_{m=1}^{M} \left[\frac{p_{mt}}{p_{ms}} \right]^{\frac{\omega_{ms}+\omega_{mt}}{2}} . \tag{4.5}$$

The Törnqvist index is usually presented and applied in its log-change form

$$\ln P_{st}^{T} = \sum_{m=1}^{M} \left(\frac{\omega_{ms} + \omega_{mt}}{2} \right) \left[\ln p_{mt} - \ln p_{ms} \right] . \tag{4.6}$$

The log-change form offers a convenient computational form. In log-change form, the Törnqvist index is a weighted average of logarithmic price changes. Furthermore, the log-change in the price of the m-th commodity, given by

$$\ln p_{mt} - \ln p_{ms} = \ln \frac{p_{mt}}{p_{ms}} \cong \left(\frac{p_{mt}}{p_{ms}} - 1 \right),$$

represents the percentage change in the price of the m-th commodity. Hence, the Törnqvist price index, in its log-change form, provides an indication of the overall growth rate in prices (inflation rate).

There are many more formulae in index number literature but we restrict our attention to these four formulae as they represent the most commonly-used formulae. We return to discuss some of the properties of these index numbers in Section 4.5, after a brief description of quantity index numbers.

4.4 Quantity Index Numbers

Two approaches can be used in measuring quantity changes. The first approach is a direct approach, where we derive a formula that measures overall quantity changes from individual commodity-specific quantity changes, measured by q_{mt}/q_{ms}. The Laspeyres, Paasche, Fisher and Törnqvist indices can be applied directly to quantity relatives. The second approach is an indirect approach, which uses the basic idea that price and quantity changes are two components that make up the value change

over the periods s and t. So, if price changes are measured directly using the formulae in the previous section, then quantity changes can be indirectly obtained after deflating the value change for the price change. This approach is discussed below in Section 4.4.2.

4.4.1 The Direct Approach

Various quantity index formulae may be defined using price index numbers, by simply interchanging prices and quantities. The formulae described above yield

$$Q_{st}^{L} = \frac{\sum_{m=1}^{M} p_{ms} q_{mt}}{\sum_{m=1}^{M} p_{ms} q_{ms}}, \qquad Q_{st}^{P} = \frac{\sum_{m=1}^{M} p_{mt} q_{mt}}{\sum_{m=1}^{M} p_{mt} q_{ms}}, \text{ and } \quad Q_{st}^{F} = \sqrt{Q_{st}^{L} \times Q_{st}^{P}}. \tag{4.7}$$

The Törnqvist quantity index, in its multiplicative and additive (log-change) forms, is given below:

$$Q_{st}^{T} = \prod_{m=1}^{M} \left[\frac{q_{mt}}{q_{ms}} \right]^{\frac{\omega_{ms} + \omega_{mt}}{2}}, \tag{4.8}$$

$$\ln Q_{st}^{T} = \sum_{m=1}^{M} \left(\frac{\omega_{ms} + \omega_{mt}}{2} \right) \left(\ln q_{mt} - \ln q_{ms} \right). \tag{4.9}$$

The Törnqvist index in equation (4.8) is the most popular index number used in measuring changes in output quantities produced and input quantities used in production over two time periods s and t. The log-change form of the Törnqvist index in equation (4.9) is the formula generally used for computational purposes. Preference for the use of the Törnqvist index formula is due to the many important economic-theoretic properties attributed to the index by Diewert (1976, 1981) and Caves, Christensen and Diewert (1982b). Economic theory issues are discussed further in the following section.

4.4.2 The Indirect Approach

The indirect approach is commonly used for purposes of quantity comparisons over time. This approach uses the basic premise that the measurement of the price and quantity changes must account for value changes.

Value change = Price change × Quantity change

$$V_{st} = P_{st} \times Q_{st}. \tag{4.10}$$

Since V_{st} is defined from data directly as the ratio of values in periods t and s, Q_{st} can be obtained as a function of P_{st}, as shown below in equation (4.1)1:

$$Q_{st} = \frac{V_{st}}{P_{st}} = \frac{\sum_{m=1}^{M} p_{mt} q_{mt}}{\sum_{m=1}^{M} p_{ms} q_{ms}} \Bigg/ P_{st} \tag{4.11}$$

$$= \frac{\sum_{m=1}^{M} p_{mt} q_{mt} \Big/ P_{st}}{\sum_{m=1}^{M} p_{ms} q_{ms}} = \frac{\text{value in period } t\text{, adjusted for price change}}{\text{value in period } s}$$

$$\therefore\ Q_{st} = \frac{\text{value in period } t \text{ (at constant prices, in period } s)}{\text{value in period } s \text{ (at prices in period } s)}.$$

The numerator in this expression corresponds to the constant price series that are commonly used in many statistical publications. Basically, this approach states that quantity indices can be obtained from ratios of values, aggregated after removing the effect of price movements over the period under consideration.

A few important features and applications of indirect quantity comparisons are discussed below.

Constant Price Aggregates and Quantity Comparison

A direct implication of equation (4.11) and the indirect approach is that value aggregates, adjusted for price changes over time, can be considered as aggregate quantities or quantities of a composite commodity. Such price deflated series are abundant in the publications of statistical agencies. Examples of such aggregates are: gross domestic product (GDP); output of sectors, such as agriculture and manufacturing; investment series; and exports and imports of good and services.

Time series and cross-section data on such aggregates are often used as data series for use in the estimation of least-squares econometric production models, stochastic frontiers, and also in DEA calculations, where it is necessary to reduce the dimensions of the output and input vectors. This means that, even if index number methodology is not used in measuring productivity changes directly, it is regularly used in creating intermediate data series.

"Self-Duality" of Formulae for Direct and Indirect Quantity Comparisons

We examine the implications of the choice of formula for price comparisons on indirect quantity comparisons. Suppose, we construct our price index numbers using the Laspeyres formula. Then, the indirect quantity index, defined in equation (4.11), can be algebraically shown to be the Paasche quantity index, defined in equation (4.7). This means that the Laspeyres price index and the Paasche quantity index form a pair that together exactly decompose the value change. In that sense, the Paasche quantity index can be considered as the dual of the Laspeyres price index.

It is also easy to show that the Paasche price index and Laspeyres quantity index together decompose the value index, and, therefore, are dual to each other. An important question that arises is: "are there self-dual index number formulae such that the same formula for price and quantity index numbers decomposes the value index exactly?" The answer to this is in the affirmative. The Fisher index for prices and the Fisher index for quantities form a dual pair. This implies that the direct quantity index obtained using the Fisher formula is identical to the indirect quantity index derived by deflating the value change by the Fisher price index. This property is sometimes referred to as the *factor reversal test.* We consider this property in the next section.

The Törnqvist index, due to the geometric nature of its formula, does not have the property of self-duality. This means that if we use the Törnqvist price index, then the indirect quantity index that is derived would be different from the quantity index derived using the Törnqvist index in equation (4.8) directly.

If the direct and indirect approaches lead to different answers, or different numerical measures of quantity changes, a problem of choice often arises as to which approach should be used in a given practical application. This problem is discussed below.

Direct versus Indirect Quantity Comparisons

Some of the analytical issues involved in a choice between direct and indirect quantity comparisons are discussed in Allen and Diewert (1981). From a practical point of view, such a choice depends on the type of data available, the variability in the price and quantity relatives, as well as the theoretical framework used in the quantity comparisons.

From a practical point of view, a researcher rarely has the luxury of choosing between direct and indirect comparisons. If the problem involves the use of aggregative data, then quantity data are usually only available in the form of constant price series. In such cases, data unavailability solves the problem[GEB2].

The second point concerns the reliability of the underlying index. Since an index number is a scalar-valued representation of changes that are observed for different commodities, the reliability of such a representation depends upon the variabilities that are observed in the price and quantity changes for the different commodities. If price changes tend to be uniform over different commodities, then the price index provides a reliable measure of the price changes. A similar conclusion can be drawn for quantity index numbers. The relative variability in the price and quantity ratios, p_{mt}/p_{ms} and q_{mt}/q_{ms} ($m = 1, 2, ...,M$) provides a useful clue as to which index is more reliable. If the price ratios exhibit less variability (relative to the quantity ratios), then an indirect quantity index is advocated, and if quantity relatives show less variability, then a direct quantity index is preferred. Variability in these ratios can be measured using standard variance measures.

Following on from this important consideration, it is worth noting that price changes over time tend to be more uniform across commodities than commodity changes. Price changes for different commodities are usually deviations from an underlying rate of overall price change. In contrast, quantity ratios tend to exhibit considerable variation across different commodities, even in the case of changes over time.

Another point of significance is that if price (quantity) ratios exhibit little variability then most index number formulae lead to very similar measures of price (quantity) change. There is more concurrence of results arising out of different formulae, and, therefore, the choice of a formula has less impact on the measure of price (quantity) change derived.

Finally, in the context of output and productivity comparisons, direct quantity comparisons may offer theoretically more meaningful indices as they utilise the constraints underlying the production technologies. Diewert (1976, 1983) and Caves, Christensen and Diewert (1982a) suggest that direct input and output quantity indices, based on the Törnqvist index formula, are theoretically superior under certain conditions. Diewert (1992) shows that the Fisher index performs very well with respect to both theoretical and test properties. An additional point in favour of the Fisher index is that it is self-dual, in that it satisfies the factor-reversal test. In addition, the Fisher index is defined using the Laspeyres and Paasche indices. Therefore, the index is easier to understand and it is also capable of handling zero quantities in the data set.

Balk (1998) suggests that under the assumption of behaviour under revenue constraints, productivity indices are best computed using indirect quantity measures. Given these results, from a theoretical point of view, the choice between direct and indirect quantity (input or output) comparisons should be based on the assumptions on the behaviour of the firm or decision making unit.

All the evidence and discussion on this issue of choice between direct and indirect quantity comparisons suggests that a decision needs to be made on pragmatic considerations as well as on pure analytical grounds.

4.5 Properties of Index Numbers: The Test Approach

In view of the existence of numerous index number formulae, Fisher (1922) proposed a number of intuitively meaningful criteria, called *tests*, to be satisfied by the formulae. These tests are used in the process of choosing a formula for purposes of constructing price and quantity index numbers. An alternative (yet closely related) framework is to state a number of properties, in the form of axioms, and then to find an index number that satisfies a given set of axioms. This approach is known as the *axiomatic approach* to index numbers. Eichorn and Voeller (1976) and Balk (1995) provide summaries of the axiomatic approach, the latter giving emphasis to price index number theory. Diewert (1992) provides a range of axioms for consideration in productivity measurement and recommends the use of the Fisher index. It is not within the scope of this book to provide details of the axiomatic approach or to delve deeply into various tests originally proposed by Fisher.[2] The main purpose of this brief section is to provide an intuitive and non-rigorous treatment of some of the tests and state two results of some importance to productivity measurement.

Let P_{st} and Q_{st} represent price and quantity index numbers, which are both real-valued functions of the prices and quantities (M commodities) observed in periods s and t, denoted by M-dimensional column vectors, $\mathbf{p}_s$, $\mathbf{p}_t$, $\mathbf{q}_s$, $\mathbf{q}_t$. Some of the basic and commonly-used axioms are listed below.

Positivity: The index (price or quantity) should be everywhere positive.

Continuity: The index is a continuous function of the prices and quantities.

Proportionality: If all prices (quantities) increase by the same proportion then P_{st} (Q_{st}) should increase by that proportion.

Commensurability or Dimensional Invariance: The price (quantity) index must be independent of the units of measurement of quantities (prices).

Time-reversal Test: For two periods s and t: $P_{st} = \dfrac{1}{P_{ts}}$.

Mean-value Test: The price (or quantity) index must lie between the respective minimum and maximum changes at the commodity level.

[2] ILO (2004), *Consumer Price Index Manual,* has detailed discussion on the test or axiomatic approach to index numbers.

Factor-reversal Test: A formula is said to satisfy this test if the *same* formula is used for direct price and quantity indices and the product of the resulting indices is equal to the value ratio.

Circularity Test (Transitivity): For any three periods, s, t and r, this test requires that: $P_{st} = P_{sr} \times P_{rt}$. That is, a direct comparison between periods s and t yields the same index as an indirect comparison through period r.

The following two results describe the properties of the Fisher and Törnqvist indices, and thus offer justification for the common use of these indices in the context of productivity measurement.

Result 4.1: The Fisher index satisfies all the properties listed above, with the exception of the circularity test (transitivity).

In fact, Diewert (1992) shows that the Fisher index satisfies many more properties. The Fisher index satisfies the factor-reversal test which guarantees a proper decomposition of value change into price and quantity changes. This justifies the label of "ideal index" attached to the Fisher formula. The factor-reversal property shows that the direct Fisher quantity index is the same as the indirect quantity index derived by deflating the value index by the Fisher price index. The Fisher index exhibits the "self-dual" property. Diewert (1976; 1981) shows that the Fisher index is *exact* and *superlative.*[3]

Result 4.2: The Törnqvist index satisfies all the tests listed above with the exception of the factor-reversal and circularity tests.

This result and other results are proved in Eichorn and Voeller (1983). Proofs of these statements are highly mathematical but the final results are quite useful. Theil (1973, 1974) shows that the Törnqvist index fails the factor-reversal test by only a small order of approximation. Failure to satisfy the factor-reversal test is not considered to be very serious as there is no necessity for the price and quantity index numbers to be *self-dual*, and no real analytical justification for the use of the same type of formula for price as well as quantity comparisons.

Fixed Base versus Chain Base Comparisons

We now briefly touch upon the issue of comparing prices, quantities and productivity over time. In the case of temporal comparisons, in particular, within the context of productivity measurement, we are usually interested in comparing each year with the previous year, and then combining annual changes in productivity to measure changes over a given period. The index constructed using this procedure

[3] These terms are explained in the next section which deals with the theoretical foundations of some of the index numbers. At this time, we note that these terms, *exact* and *superlative*, refer to properties related to the economic theory underlying price and quantity index number measurement.

is known as a *chain index.* To facilitate a formal definition, let $I(t, t+1)$ define an index of interest for period $t+1$ with t as the base period. The index can be applied to a time series with $t = 0, 1, 2, .., T$. Then, a comparison between period t and a fixed base period, 0, can be made using the following chained index of comparisons for consecutive periods:

$$I(0, t) = I(0, 1) \times I(1, 2) \times ... \times I(t-1, t)$$

As an alternative to the chain-base index, it is possible to compare period 0 with period t using any one of the formulae described earlier. The resulting index is known as the *fixed-base index.*

Most national statistical offices make use of a fixed-base Laspeyres index mainly because the weights remain the same for all the fixed-base index computations. Usually, the base periods are shifted on a regular basis.

There is a considerable index number literature focusing on the relative merits of fixed- and chain-base indices. A good survey of the various issues can be found in Forsyth (1978), Forsyth and Fowler (1981) and Szulc (1983). From a practical angle, especially with respect to productivity measurement, a chain index is more suitable than a fixed-base index. Since the chain index involves only comparisons with consecutive periods, the index is measuring smaller changes. Therefore, some of the approximations involved in the derivation of theoretically meaningful indices are more likely to hold. Another advantage is that comparisons over consecutive periods mean that the Laspeyres-Paasche spread is likely to be small indicating that most index number formulae result in indices which are very similar in magnitude. The only drawback associated with the chain index is that the weights used in the indices need to be revised every year.

The use of a chained index does not result in transitive index numbers. Even though transitivity is not considered essential for temporal comparisons, it is necessary in the context of multilateral comparisons. The question of transitivity in multilateral comparisons is considered in some detail in Section 4.7, as multilateral TFP studies are now quite common.

Which Formula to Choose?

The foregoing discussion indicates that the choice of formula is essentially between the Fisher and Törnqvist indices. Both of these indices possess important properties and satisfy a number of axioms. If published aggregated data are used then it is necessary to check what formula was used in creating the series. It is very likely that Laspeyres or Paasche indices are used in such data series. If the indices are being computed for periods which are not far apart then differences in the numerical values of the Fisher and the Törnqvist indices are likely to be quite minimal. Further, both of these indices also have important theoretical properties. While, in practice, the Törnqvist index seems to be preferred, use of the Fisher index is

recommended due to its additional self-dual property and its ability to accommodate zeros in the data.

4.6 The Economic-Theoretic Approach

This[GEB3] chapter is primarily devoted to a detailed examination of the economic-theoretic foundations of the various index numbers discussed in Section 4.3. Economic theory is important in understanding what these index numbers actually measure and make proper applications of them. This section is also important in that it provides a base from which we integrate the three principal approaches, namely, the index number, DEA and SFA approaches, in the context of productivity and efficiency measurement.

The economic-theoretic approach to index numbers is also known as the functional approach to index numbers, since the approach postulates a functional relationship between observed prices and quantities for inputs as well as outputs. In the case of productivity measurement, the economic theory relevant to production (i.e., the microeconomic theory of the firm) is even more relevant. The functional approach contrasts with the simple mathematical approach, usually known as the test (or axiomatic) approach that is considered in earlier sections, which relies on a range of well-defined properties, tests or axioms.

Before embarking on the economic-theoretic approach, we formally establish some of the tools used to derive the various index numbers of interest. We consider the general case involving M outputs and N inputs. Let s and t represent two time periods or firms; p_{ms} and p_{mt} represent output prices for the m-th commodity in periods s and t, respectively; q_{ms} and q_{mt} represent output quantities in periods s and t, respectively ($m = 1,2,...,M$); w_{ms} and w_{mt} represent input prices in periods s and t, respectively; and x_{ns} and x_{nt} represent input quantities in periods s and t, respectively ($n = 1,2,...,N$). Further, let $\mathbf{p}_t$, $\mathbf{p}_s$, $\mathbf{q}_t$, $\mathbf{q}_s$, $\mathbf{w}_s$, $\mathbf{w}_t$, $\mathbf{x}_t$ and $\mathbf{x}_s$ represent vectors of non-negative real numbers of appropriate dimensions. Let S^s and S^t represent the production technologies in periods s and t, respectively.[4][GEB4] In deriving various price and quantity (for inputs and outputs) index numbers we make use of revenue and cost functions, and the input and output distance functions discussed in Chapter 3.

The approach we examine revolves around two basic indices proposed decades ago. All price index numbers, input as well as output price indices, are based on the famous Konus (1924) index and the approach discussed in Fisher and Shell (1972). The Konus index, in its original form, provided an analytical framework for measuring changes in consumer prices, i.e., the construction of cost of living indices.

[4] In the presence of multiple outputs, it is not possible to express the production technology in the form of a simple production function since the production function is a "real-valued function" showing the maximum level of output that can be produced with a given level of inputs. As a result, we use sets to represent the technology. Alternatively, we could use a transformation function.

The input and output quantity index numbers, and productivity indices, are all based on the ideas of Malmquist and the distance function approach, outlined in Malmquist (1953). No treatment of these aspects is complete without reference to Diewert (1976, 1978, 1981), Caves, Christensen and Diewert (1982*a* and 1982*b*), Färe, Grosskopf and Lovell (1985, 1994), Färe and Primont (1995) and Balk (1998).

The results presented in the ensuing sections are drawn from Caves, Christensen and Diewert (1982*b*), Diewert (1983, 1992), Färe and Primont (1995), Färe, Grosskopf and Roos (1998) and Balk (1998). The foundations for some of this work, in fact, dates back to Diewert (1976) where the ideas of *exact* and *superlative* index numbers were first introduced. The exposition presented here is not exhaustive in its coverage, rather it is to provides a flavour of what is on offer.

An important point to keep in mind is that *the economic-theoretic approach to index numbers assumes that the firms observed in periods s and t are both technically and allocatively efficient.*[5]. This means that observed output and input data are assumed to represent optimising behaviour involving revenue maximisation and cost minimisation, or, in some cases, constrained optimisation involving revenue maximisation with cost constraints, etc. A compensating factor in lieu of these somewhat restrictive assumptions is that we are able to derive various index numbers of interest with just two observations on prices and quantities for the two periods, *s* and *t*. We now examine different classes of index numbers derived using this approach.

4.6.1 Output Price Indices

For a given level of inputs, **x**, let the (maximum) revenue function[6] be[GEB5] defined, for technology in period-t, as

$$R^t(\mathbf{p},\mathbf{x}) = \max_{\mathbf{q}} \{\mathbf{pq}: (\mathbf{x},\mathbf{q}) \text{ is feasible in } S^t\} \tag{4.12}$$

We can illustrate this function by using the production possibility curve and isorevenue line (for the case of two outputs), as in Figure 4.1. The point of tangency between the production possibility curve and the isorevenue line indicates the combination of the two outputs (q_1 and q_2) that maximise revenue, given the input vector, **x**, the output price vector, $\mathbf{p}_t$, and the technology, S^t.

[5] Balk (1998) relaxes the assumption of technical efficiency in deriving various index numbers. However, for purposes of this section, we maintain the assumption of allocative efficiency.
[6] Refer to Chapter 2 for the definition and properties of the *revenue function*.

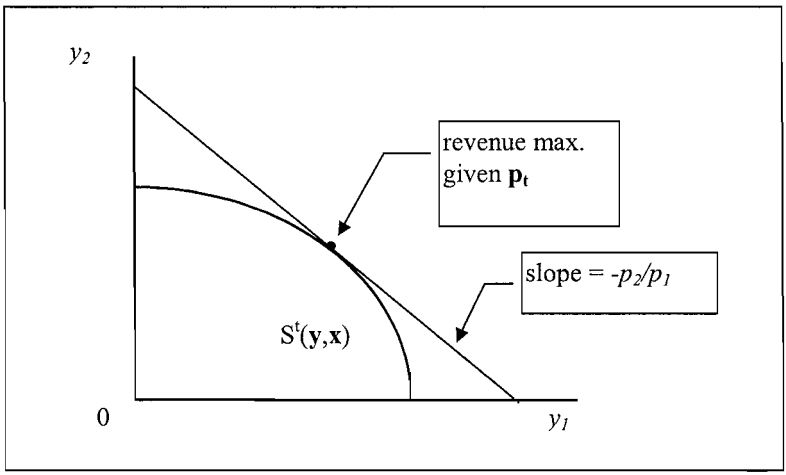

Figure 4.1 Revenue Maximisation

The output price index, due to Fisher and Shell (1972) and Diewert (1980), based on period-t technology, is defined as

$$P_o^{\,t}(\mathbf{p}_s,\mathbf{p}_t,\mathbf{x}) = \frac{R^t(\mathbf{p}_t,\mathbf{x})}{R^t(\mathbf{p}_s,\mathbf{x})}. \tag{4.13}$$

This index is the ratio of the maximum revenues possible with the two price vectors, $\mathbf{p}_s$ and $\mathbf{p}_t$, using a fixed level of inputs, $\mathbf{x}$, and period-t technology. This is illustrated in Figure 4.2, where we observe the revenue maximising points associated with the price vectors, $\mathbf{p}_t$ and $\mathbf{p}_s$.

The output price index in equation (4.13) can also be defined using period-s technology leading to

$$P_o^{\,s}(\mathbf{p}_s,\mathbf{p}_t,\mathbf{x}) = \frac{R^s(\mathbf{p}_t,\mathbf{x})}{R^s(\mathbf{p}_s,\mathbf{x})}. \tag{4.14}$$

Some basic features of the two output price index numbers in equations (4.13) and (4.14) can be noted. These indices depend upon the state of technology, whether it is the period-t or period-s (or, for that matter, any other period) technology involved, and then on the input vector, $\mathbf{x}$, at which the index is calculated. It is useful to ask the question: under what conditions are these indices independent of these two factors?

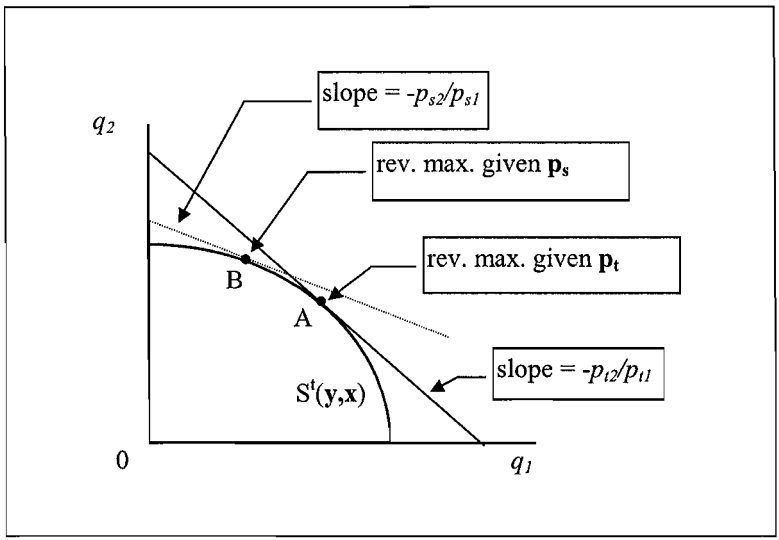

Figure 4.2 Output Price Index

These indices are independent of **x** if and only if the technology is output homothetic. Following Färe and Primont (1995, p. 68), a production technology is output homothetic if the output sets P(**x**) depend upon the output set for the unit input vector (input quantities equal to one for all inputs) and a real-valued function, G(**x**), of **x**. In simple terms, the production possibility curves in Figures 4.1 and 4.2, for different input vectors, **x**, are all parallel shifts of the production possibility curve for the unit-input vector.

In a similar vein, it can be shown that if the technology exhibits implicit output neutrality then the indices are independent of which period's technology is used in the derivation.

The output price index numbers in equations (4.13) and (4.14) satisfy a number of standard properties. For example, these indices satisfy monotonicity, linear homogeneity, identity, proportionality, independence of units of measurement, transitivity for fixed t and **x**, and time-reversal properties.[7]

Since $\mathbf{x}_t$ and $\mathbf{x}_s$ are the actual input levels used in periods t and s, we can define the indices in equations (4.13) and (4.14) using the actual input levels, leading to two natural output price index numbers:

[7] Most of these properties can be established using simple logic and, in some cases, using the properties of the revenue function that flow from the basic assumptions on the production technology discussed in Section 3.2.

$$P_o^t(\mathbf{p}_s,\mathbf{p}_t,\mathbf{x}_t)=\frac{R^t(\mathbf{p}_t,\mathbf{x}_t)}{R^t(\mathbf{p}_s,\mathbf{x}_t)} \tag{4.15}$$

$$P_o^s(\mathbf{p}_s,\mathbf{p}_t,\mathbf{x}_t)=\frac{R^s(\mathbf{p}_t,\mathbf{x}_t)}{R^s(\mathbf{p}_s,\mathbf{x}_t)} \tag{4.16}$$

While these indices have some intuitive appeal, computing them requires the knowledge of the functional form of the revenue functions as well as the numerical values of the parameters underlying the functions. In essence, we need to have a complete description of the technology, which is a hopeless task since we are trying to measure these changes based only on observed prices and quantities in these two periods. The following results illustrate how we can get close to the theoretically defined index numbers in equations (4.15) and (4.16).

Result 4.3: Under the assumptions of optimal behaviour (allocative and technical efficiency) and regularity conditions on the production technologies, the two index numbers in equations (4.15) and (4.16) are, respectively, bounded by the Laspeyres and Paasche indices, namely,

$$\text{Laspeyres Price Index} = \frac{\sum_{m=1}^{M} p_{mt}q_{ms}}{\sum_{m=1}^{M} p_{ms}q_{ms}} \le P_o^s(\mathbf{p}_s,\mathbf{p}_t,\mathbf{x}_s)$$

$$\text{Paasche Price Index} = \frac{\sum_{m=1}^{M} p_{mt}q_{mt}}{\sum_{m=1}^{M} p_{ms}q_{mt}} \ge P_o^t(\mathbf{p}_s,\mathbf{p}_t,\mathbf{x}_s).$$

Proofs of these results are obtained by recalling that we have assumed that $\mathbf{q}_t$ is optimal for $\mathbf{x}_t$ at price $\mathbf{q}_t$ and $\mathbf{q}_s$ is optimal for $\mathbf{x}_s$ at price $\mathbf{p}_s$. For example, in the case of the Paasche index, this implies that the revenue produced at the point A in Figure 4.2 must be equal to $\Sigma_m p_{mt}q_{mt}$ and we also know that the revenue produced at the point B in Figure 4.2 must be at least as large as $\Sigma_m p_{ms}q_{mt}$ (because that revenue is associated with a y-vector that produces maximum revenue for the given $\mathbf{p}_s$). Hence, we find that the Paasche index is at least as large as $P^t(\mathbf{p}_s,\mathbf{p}_t,\mathbf{x}_s)$. A similar argument is involved for the Laspeyres index.

The main point to note from Result 4.3 is that the Laspeyres and Paasche indices, which were defined on a heuristic basis (within the atomistic approach), provide lower and upper bounds for theoretical output price indices defined using production technologies and optimising behaviour. The following two results show that even though the two indices in equations (4.15) and (4.16) cannot be individually

determined, the geometric mean of these two indices can be reasonably well approximated.

Result 4.4: A reasonable approximation to the geometric mean of the two indices in equations (4.15) and (4.16) is provided by the Fisher output price index number,

$$[\, P_o^t(\mathbf{p}_s, \mathbf{p}_t, \mathbf{x}_t) \times P_o^s(\mathbf{p}_s, \mathbf{p}_t, \mathbf{x}_s)\,]^{1/2} \cong \left[\frac{\sum_{m=1}^{M} p_{mt} q_{ms}}{\sum_{m=1}^{M} p_{ms} q_{ms}} \times \frac{\sum_{m=1}^{M} p_{mt} y_{mt}}{\sum_{m=1}^{M} p_{ms} y_{mt}} \right]^{0.5}$$

$$= \text{Fisher Price Index.}$$

The accuracy of this approximation relies on how symmetric the Laspeyres and Paasche index numbers are, relative to the respective economic-theoretic index numbers in equations (4.15) and (4.16).

Suppose we now assume that the revenue functions have the translog form.[8] This assumption is essentially in line with the fact that the translog function is a flexible form and provides a second-order approximation to the unknown revenue function.[9] However, alternative specifications, such as the normalised quadratic, generalised Leontief and the generalised McFadden functions, are also available. Diewert and Wales (1987) discuss the properties of various flexible functional forms.

The translog revenue function is given by[10]

$$\ln R^t(\mathbf{x},\mathbf{p}) = \alpha_{0t} + \sum_{k=1}^{K} \alpha_{kt} \ln x_k + \sum_{m=1}^{M} \beta_{mt} \ln p_m + \tfrac{1}{2} \sum_{k=1}^{K} \sum_{j=1}^{K} \alpha_{kjt} \ln x_k \ln x_j$$
$$+ \tfrac{1}{2} \sum_{m=1}^{M} \sum_{j=1}^{M} \beta_{mjt} \ln p_m \ln p_j + \tfrac{1}{2} \sum_{k=1}^{K} \sum_{m=1}^{M} \gamma_{kmt} \ln x_k \ln p_m$$

where $\alpha_{kjt} = \alpha_{jkt}$ and $\beta_{mjt} = \beta_{jmt}$ and $\gamma_{kmt} = \gamma_{mkt}$. The following result uses a translog specification for the revenue function.

Result 4.5: If the revenue functions for periods s and t are represented by translog functions, with second-order coefficients being equal for periods s and t ($\alpha_{kjt} = \alpha_{kjs}$,

[8] Examples of translog functions are given in Chapter 2 and econometric estimation of translog productions is discussed in Chapter 8.

[9] The same translog functional form can be used in approximating any unknown function. We also make use of the translog form to approximate input cost functions, input and output distance functions as well as profit functions. In each case, we use different variables depending on the function involved.

[10] Note that some restrictions on the β- and γ-parameters need to be imposed to make the function satisfy linear homogeneity in output prices. We do not provide detailed translog specifications for each of the functions we consider in this chapter.

$\beta_{mjt} = \beta_{mjs}$,, $\gamma_{kmt} = \gamma_{kms}$) then the geometric mean of the two price indices in equations (4.15) and (4.16) is equal to the Törnqvist output price index

$$\left[P_o^{\,s}(\mathbf{p}_s,\mathbf{p}_t,\mathbf{x}_s)\times P_o^{\,t}(\mathbf{p}_s,\mathbf{p}_t,\mathbf{x}_t)\right]^{0.5} = \prod_{m=1}^{M}\left(\frac{p_{mt}}{p_{ms}}\right)^{\frac{r_{mt}+r_{ms}}{2}}, \tag{4.17}$$

where $r_{mt} = \dfrac{p_{mt}q_{mt}}{\sum_{m=1}^{M} p_{mt}q_{mt}}$ and $r_{ms} = \dfrac{p_{ms}q_{ms}}{\sum_{m=1}^{M} p_{ms}q_{ms}}$ are the value shares in periods s and t, respectively. A proof of this result is in Diewert (1983).

Some comments on these results are in order. The importance of this result is that, even though the theoretical indices in equations (4.15) and (4.16) require the knowledge of the parameters of the revenue function, their geometric mean is equal to the Törnqvist index and the index can be computed from the observed price and quantity data. Knowledge of the parameters of the translog functions is unnecessary.

If the translog function is replaced by a quadratic function then the Fisher index can be shown to be equal to the geometric mean on the left-hand side of equation (4.17) in Result 4.5 (see Diewert (1992) for the proof).

The Törnqvist index is considered to be *exact* for the translog revenue function, and it is considered *superlative* since the translog function is a flexible functional form (i.e., provides a second-order approximation to any arbitrary function). The Fisher index is *exact* for a quadratic function and, hence, is also *superlative*. A more detailed exposition of exact and superlative indices is available in Diewert (1976).

4.6.2 Input Price Indices

The following framework for input price index numbers is essentially adapted from the Konus (1924) cost-of-living index which measures the changes in the cost of maintaining certain utility levels at different sets of prices. Extending this concept we can measure input price index numbers by comparing costs of producing a certain vector of outputs, given different input price vectors. In this process we need to define a cost function, associated with a given production technology, for a given output level, $\mathbf{q}$, namely

$$C^t(\mathbf{w},\mathbf{q}) = \min_{\mathbf{x}}\left\{\mathbf{wx} \mid \mathbf{x},\mathbf{q}\in S^t\right\}. \tag{4.18}$$

The cost function, $C^t(\mathbf{w},\mathbf{q})$, is the minimum cost of producing $\mathbf{q}$, given period-t technology, using input price vector, $\mathbf{w}$.

It is easy to check that the cost function, $C^t(\mathbf{w},\mathbf{q})$, is linearly homogeneous in w[11] and it is non-decreasing in the output vector, **q**. We can use the cost function to define input price index numbers. Given the input prices, $\mathbf{w}_t$ and $\mathbf{w}_s$, in periods t and s, we can define the input price index as the ratio of the minimum costs of producing a given output vector **q** using an arbitrarily selected production technology, S^j (j = s,t). Then the index is given by

$$P_i^j(\mathbf{w}_s,\mathbf{w}_t,\mathbf{q})=\frac{C^j(\mathbf{w}_t,\mathbf{q}|S)}{C^j(\mathbf{w}_s,\mathbf{q}\{S)}, \tag{4.19}$$

where the cost functions are defined using technology set, S. The cost elements in equation 4.19 can be seen from Figure 4.3 below. Let the isoquant under technology, S^s, for a given output level, **q**, be represented by Isoq(**q**)-S^s. The two sets of input prices, $\mathbf{w}_s$ and $\mathbf{w}_t$ can be represented by isocost lines AA′ and BB′, respectively. Minimum-cost combinations of inputs, producing output vector, **q**, for these two input price vectors are given by the points, **x*** and **x****. These points are obtained by shifting lines AA′ and BB′ to aa′ and bb′, respectively, where they are tangential to Isoq(y)-S^s. The input price index number in equation 4.19 for this two input case is then given by the ratio of the costs at points, **x*** and **x****.

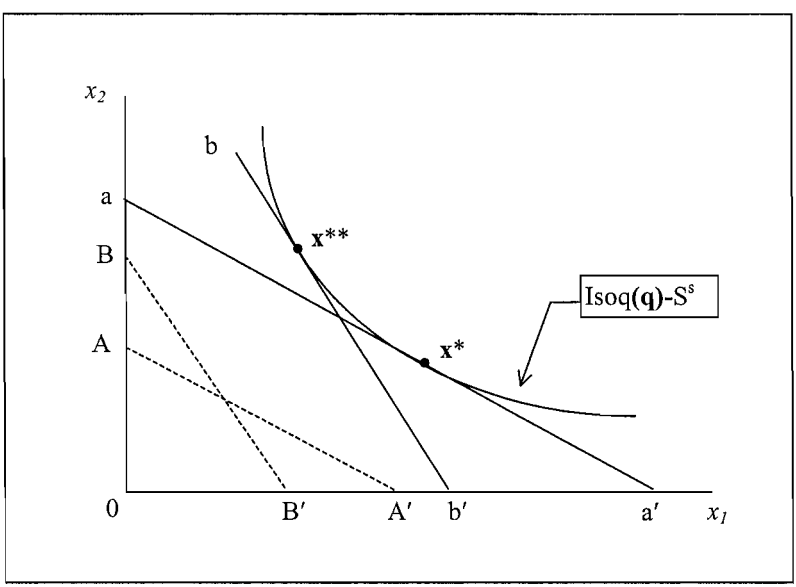

Figure 4.3 Input Price Index

[11] This implies that if all the input prices are multiplied by a common factor, λ, then the cost associated with the new prices is λ times the original cost. See Chapter 2 for other properties of the cost function.

The input price index defined above can be shown to satisfy many useful properties, including: monotonicity, linear homogeneity in input prices, independent of units of measurement, proportionality and transitivity (for a fixed **q** and technology). These properties follow from duality theory and the fact that the cost function, $C^j(\mathbf{w},\mathbf{q})$, is associated with a production technology, S^j, with certain characteristics.[12]

In order to be able to actually compute the input price index in equation 4.19 we need to specify the technology and also the output level, **q**, at which we wish to compute the index. Two points can be made here. First, the price index is independent of which period technology we use if and only if the technology exhibits implicit Hicks input neutrality. Second, the index $P_i(\mathbf{w}_t, \mathbf{w}_s, \mathbf{q})$, for a given technology, is independent of the output level, **q**, if and only if the technology exhibits input homotheticity (see Färe and Primont, 1995).

If the technology does not satisfy these conditions, then we can define many input price index numbers using alternative specifications for technology, S, and the output vector, **q**. Two natural specifications are to use the period-*s* and period-*t* technologies, along with the output vectors, $\mathbf{q}_s$ and $\mathbf{q}_t$. These result in the following input price index numbers[13]

$$P_i^s(\mathbf{w}_s,\mathbf{w}_t,\mathbf{q}_s)=\frac{C^s(\mathbf{w}_t,\mathbf{q}_s)}{C^s(\mathbf{w}_s,\mathbf{q}_s)}, \tag{4.20}$$

and

$$P_i^t(\mathbf{w}_s,\mathbf{w}_t,\mathbf{q}_t)=\frac{C^t(\mathbf{w}_t,\mathbf{q}_t)}{C^t(w_s,\mathbf{q}_t)}. \tag{4.21}$$

Observe that under the assumptions of allocative and technical efficiency, the observed input costs, $\mathbf{w}_s\mathbf{x}_s$ and $\mathbf{w}_t\mathbf{x}_t$, are equal to $C^s(\mathbf{w}_s,\mathbf{q}_s)$ and $C^t(\mathbf{w}_t,\mathbf{q}_t)$, respectively. We now state the following two results without proofs. Proof of Result 4.6 is fairly straightforward, while the proof of Result 4.7 is a bit more involved.

Result 4.6: Under our assumptions on the production technologies in periods, t and s, and given the optimal behaviour of the firm in these periods, the Laspeyres and Paasche indices provide upper and lower bounds to the economic-theoretic index numbers in equations 4.20 and 4.21. Also the geometric mean of these indices can be approximated by the Fisher price index numbers for input prices.

[12] Refer to the section on duality in Chapter 2 for a discussion of these properties.

[13] These indices correspond to the usual Laspeyres and Paasche-type index numbers because they rely on the base- and current-period technologies and output vectors.

The next result is based on the assumption that the cost function, $C(\mathbf{w},\mathbf{q})$, has a translog form,[14] along with appropriate restrictions on the parameters of the cost function to ensure linear homogeneity in input prices.

Result 4.7: If the technologies in periods t and s are represented by the translog cost function, with the additional assumption that the second order coefficients are identical in these periods, then, under the assumption of technical and allocative efficiency the geometric mean of the two input price index numbers in equations 4.20 and 4.21 is given by the Törnqvist price index number applied to input prices and quantities. That is

$$\left[P_i^s(\mathbf{w}_s,\mathbf{w}_t,\mathbf{q}_s)\times P_i^t(\mathbf{w}_s,\mathbf{w}_t,\mathbf{q}_t)\right]^{1/2} = \prod_{n=1}^{N}\left(\frac{w_{nt}}{w_{ns}}\right)^{\frac{s_{nt}+s_{ns}}{2}}, \tag{4.22}$$

where s_{nt} and s_{nt} are the input expenditure shares of n-th input in periods t and s, respectively. The right-hand side of equation 4.22 is the Törnqvist price index defined in Section 4.3.

Results 4.6 and 4.7 imply that the Fisher and Törnqvist indices, discussed in Section 4.3, can be applied in measuring changes in input prices and at the same time have a proper economic-theoretic framework to support their use. These results also illustrate that, under certain assumptions, it is not necessary to know the numerical values of the parameters of the cost or revenue function or the underlying production technology: it is sufficient to have the observed input price and quantity data to measure changes in input prices.

Result 4.7 shows that the Törnqvist input price index is *exact* for the geometric mean of the two theoretical indices, when the underlying cost function is translog and hence can also be considered *superlative*. Diewert (1983) shows that the Fisher input price index, while it provides an approximation as in Result 5.4, is also *exact* for a quadratic cost function. Diewert (1992) introduces yet another specification for the cost function under which the Fisher input price index can be shown to be exact.

An important point to note here is that much of the recent emphasis and popularity enjoyed by the Törnqvist and Fisher indices owe a great deal to the work of Diewert (1976, 1981, 1983 and 1992) and Caves, Christensen and Diewert (1982b). However, it is important to be aware of the assumptions underlying these results.

[14] Note here that we made a similar assumption in the context of output price index number where the *revenue function* was assumed to be a translog form. However, due to the non-dual nature of the translog function, one assumption does not imply the other. The translog functional form is assumed for input and output distance functions that are used in the next two subsections.

4.6.3 Output Quantity Indices

This section deals with the theoretical approach used in deriving output quantity index numbers. It is at this point, we make use of the distance functions introduced in Chapter 3. Use of distance functions in defining quantity (both output and input) index numbers is based on the work of Malmquist (1953).

Unlike the case of price index numbers, three possible strategies can be followed in deriving theoretically sound quantity index numbers. These are discussed below. We describe the first two approaches only briefly and then focus mainly on the Malmquist index defined using the distance function.

The Method of Deflation

This approach is discussed in Fisher and Shell (1972) and follows closely the indirect quantity index numbers discussed earlier. The approach here is to divide the value index by the output price index. Using period-t technology, at input level x_t, the output quantity index is given by

$$Q_o^t(\mathbf{p}_s, \mathbf{p}_t, \mathbf{x}_t, \mathbf{q}_s, \mathbf{q}_t) = \frac{\sum_{m=1}^{M} p_{mt} q_{mt}}{\sum_{m=1}^{M} p_{ms} y_{ms}} \Bigg/ P^t(\mathbf{p}_s, \mathbf{p}_t, \mathbf{x}_t) \tag{4.23}$$

This equation provides a quantity index which is a generalisation of equation 4.11, obtained by deflating the value index by a theoretically meaningful index of output prices. Equation 4.23 makes use of period-t technology and input vector, $\mathbf{x}_t$. An index similar to equation 4.23 can be defined using period-s technology and input vector, $\mathbf{x}_s$.

It is necessary to make a choice about which formula we use in measuring the output price changes. Either Törnqvist or Fisher output price index numbers could be used in equation 4.23. As we observed in Section 4.6.2, the final choice depends upon the functional form of the revenue function we are prepared to assume.

The Samuelson and Swamy Approach

In this section, we briefly describe the approach suggested by Samuelson and Swamy (1974) for measuring changes in output levels. Their approach uses the revenue function, R(**x**,**p**), associated with an output price vector, **p**, and input vector, **x**, under a given production technology.

Under this approach the quantity index is defined as the ratio of the revenue functions derived from the inputs $\mathbf{x}_s$ and $\mathbf{x}_t$, in periods s and t, at some arbitrarily defined price vector, **p**, namely

$$
\begin{aligned}
Q_o^{SS}(\mathbf{x}_s, \mathbf{x}_t, \mathbf{p}) &= \frac{R^t(\mathbf{p}, \mathbf{x}_t)}{R^s(\mathbf{p}, \mathbf{x}_s)} \\
&= \frac{R^t(\mathbf{p}, \mathbf{x}_t)}{R^s(\mathbf{p}, \mathbf{x}_t)} \times \frac{R^s(\mathbf{p}, \mathbf{x}_t)}{R^s(\mathbf{p}, \mathbf{x}_s)}.
\end{aligned}
\tag{4.24}
$$

It is possible to interpret the first ratio on the right hand-side of the last line of equation 4.24 as a measure of the technical change from periods to period t, and the second component as a measure of output change, as measured from revenue changes, due to change in input use under the period-s technology.

The Malmquist Approach

The Malmquist approach is the most commonly used approach for output comparisons. Using the distance functions defined in Section 3.2, the Malmquist output index, based on technology in period-t, is defined as

$$
Q_o^t(\mathbf{q}_s, \mathbf{q}_t, \mathbf{x}) = \frac{d_o^t(\mathbf{x}, \mathbf{q}_t)}{d_o^t(\mathbf{x}, \mathbf{q}_s)},
\tag{4.25}
$$

for an arbitrarily selected input vector, $\mathbf{x}$.

A similar Malmquist index can be defined using period-s technology. In fact, we can also define many alternative indices using different levels of $\mathbf{x}$. As before, the index defined in equation 4.25 is independent of the technology involved, if and only if the technology exhibits Hicks output neutrality. The quantity index is independent of the input level, $\mathbf{x}$, if and only if the technology is output homothetic. Even in the cases where these assumptions hold, we still need to know the functional form of the distance function as well as numerical values of all the parameters involved. The index number approach attempts to bypass this problem by providing approximations to the index in equation 4.25 when we are not sure of the functional form, or do not have adequate information to estimate the parameters of the distance function, even when we know the form of the function.[15]

If we consider output quantity indices based on technology in periods s and t, along with the inputs used in these periods, we have two possible measures of output change, given by $Q_o^s(\mathbf{q}_t, \mathbf{q}_s, \mathbf{x}_s)$ and $Q_o^t(\mathbf{q}_t, \mathbf{q}_s, \mathbf{x}_t)$. There are many standard results

[15] Such estimation requires considerable panel data. These issues are further considered in the discussion of data envelopment analysis and stochastic frontiers in the following chapters. At this point, the strength of the index number approach lies in the fact that we can measure output changes without a lot of data.

of interest[16] that relate these indices to the standard Laspeyres and Paasche quantity index numbers, defined in Section 4.3. A result of particular interest is that the Fisher index provides an approximation to the geometric average of these two indices (see Diewert, 1981, 1983; and Balk, 1997). The following result establishes the economic-theoretic properties of the Törnqvist output index and shows why the index is considered to be an *exact* and a *superlative* index.

Result 4.8: If the distance functions for periods s and t are both represented by translog functions with identical second-order parameters, then a geometric average of the Malmquist output indices in equation 4.25, based on technologies of periods s and t, with corresponding input vectors $\mathbf{x}_s$ and $\mathbf{x}_t$, is equivalent to the Törnqvist output quantity index. That is,

$$\left[Q_o^s(\mathbf{q}_s,\mathbf{q}_t,\mathbf{x}_s)\times Q_o^t(\mathbf{q}_s,\mathbf{q}_t,\mathbf{x}_t)\right]^{1/2} = \text{Törnqvist output index.} \tag{4.26}$$

This result implies that the Törnqvist index is *exact* for the geometric mean of the period-t and period-s theoretical output index numbers when the technology is represented by a translog output distance function. Since the translog functional form is *flexible* (i.e., it provides a second-order approximation to an arbitrary, twice continuously differentiable, functional form), the Törnqvist index is also consider to be *superlative*.

If the translog functions are replaced by quadratic functions, with appropriate normalisation and restrictions, to represent the output distance function, then the left-hand side of equation 4.26 can be shown to be equal to the Fisher output quantity index number, which in turn establishes the *exact* and *superlative* nature of the Fisher output quantity index.

It should also be noted that if the Laspeyres or Paasche indices are used to calculate output (or input) indices, then this would imply an underlying linear technology for the production structure. If we cast our minds back to the "S-shaped" production functions, discussed in Chapter 1, it is obvious that a linear technology would be a simplifying assumption. It would imply constant returns to scale and constant marginal products throughout, together with other restrictive properties.

In terms of the choice between the three alternative measures of quantity change discussed in this subsection, the Malmquist index is the only index that satisfies the homogeneity property which states that, if $\mathbf{q}_t=\lambda\mathbf{q}_s$, then $Q_o^t(q_s,q_t,\mathbf{x})=Q_o^t(q_s,\lambda q_s,x)=\lambda$. This property does not hold for the Fisher-Shell and Samuelson-Swamy approaches. Diewert (1983) states some necessary and sufficient conditions under which all the three approaches are equivalent. Mainly these conditions revolve around

[16] These results are very similar to Results 4.3, 4.4 and 4.5, stated in the context of output price index numbers, and Results 4.6 and 4/7, stated for input price index numbers. It is easy to restate these results for output quantity index numbers incorporating appropriate modifications.

homotheticity. Some practical issues arising out of practical applications of direct versus implicit quantity index numbers are discussed in Allen and Diewert (1981). These practical issues were discussed in Section 4.2.[17]

4.6.4 Input Quantity Indices

We now turn to the measurement of change in the input use by a firm over two time periods, t and s. An obvious strategy, which we are not going to pursue any further, is to measure input change by deflating the change in expenditure on inputs over periods s and t, by the input price index number defined earlier. In this section we describe the input quantity index number, derived using the Malmquist distance measure.

Using the concept of the input distance function, we can now define the input quantity index. Along the same lines as the output index, we can compare the levels of input vectors $\mathbf{x}_t$ and $\mathbf{x}_s$, by measuring their respective distances from a given output vector, for a given state of the production technology.

The input quantity index, based on the Malmquist input distance function, is defined for input vectors, $\mathbf{x}_s$ and $\mathbf{x}_t$, with base period-s and using period-t technology, is given by

$$Q_i^t(\mathbf{x}_s, \mathbf{x}_t, \mathbf{q}) = \frac{d_i^t(\mathbf{x}_t, \mathbf{q})}{d_i^t(\mathbf{x}_s, \mathbf{q})}. \tag{4.27}$$

As in the case of other index numbers we have defined before, the input quantity index in equation 4.27 satisfies monotonocity, linear homogeneity in the input vector $\mathbf{x}_t$, and it is invariant to scalar multiplication of the input vectors, i.e., if both input vectors are multiplied by a constant, κ, then the index remains unchanged. In addition, the index is independent of units of measurement. Proofs of these results follow from the properties of the input distance function, which can be derived using the axioms of the production technology.

Following the same approach as in the previous sections, we note that the input quantity index depends upon the output level, $\mathbf{q}$, we choose, as well as the production technology. If we use period-s technology in defining the input distance functions, then we get the following index

[17] Balk (1998) recommends the use of Fisher-Shell approach when productivity measures are derived, given the assumption of economic behaviour of the firm under cost or revenue restrictions. These results suggest that quantity index numbers derived under constrained optimisation behaviour are in the form of values deflated by the respective price index numbers.

$$Q_i^s(\mathbf{x}_s, \mathbf{x}_t, \mathbf{q}) = \frac{d_i^s(\mathbf{x}_t, \mathbf{q})}{d_i^s(\mathbf{x}_s, \mathbf{q})}. \tag{4.28}$$

The input indices defined in equations 4.27 and 4.28 are usually different. These two coincide and, in fact, are independent of which technology we choose, if the technologies in these periods exhibit implicit Hicks input neutrality (see Färe and Primont, 1995). These indices are independent of the reference output vector, **q**, used in the definitions above if and only if the technology exhibits input homotheticity.

Our main purpose is to relate this Malmquist input quantity index number to the input index number derived using some of the formulae in Section 4.3. The input index in equation 4.28, defined using base period-s technology is bounded from above by the Laspeyres quantity index. Further, the index in equation 4.27, defined on current period-t technology, is bounded from below by the Paasche quantity index. Therefore, the Fisher input quantity index provides an approximation to the geometric mean of the indices, $Q_i^s(\mathbf{x}_t,\mathbf{x}_s,\mathbf{q})$ and $Q_i^t(\mathbf{x}_t,\mathbf{x}_s,\mathbf{q})$.

If we wish to go further we need to know the functional form of the input distance functions in equations 4.27 and 4.28. If we assume a quadratic function, then the Fisher input quantity index can be shown to be equal to the geometric average of the two indices. However, if the distances functions are of the translog form,[18] if the distances functions in periods t and s have identical second-order coefficients satisfying the usual restrictions on the parameters of the translog form, and if the assumption of allocative and technical efficiency holds, then

$$\begin{aligned}\left[Q_i^s(\mathbf{x}_s, \mathbf{x}_t, \mathbf{q}) \times Q_o^t(\mathbf{x}_s, \mathbf{x}_t, \mathbf{q})\right]^{1/2} &= \text{Törnqvist Input quantity index} \\ &= \prod_{n=1}^{N} \left[\frac{x_{nt}}{x_{ns}}\right]^{(\omega_{nt}+\omega_{ns})/2}, \end{aligned} \tag{4.29}$$

where s_{nt} and s_{ns} are input cost-shares in periods t and s, respectively. This result shows that the Törnqvist index is *exact* and *superlative* for the geometric mean of Malmquist input index numbers based on the technologies of periods t and s.

In this section, we have concluded our discussion of the economic-theoretic framework for index numbers measuring changes in prices and quantities. We now turn our attention to the measurement of productivity and the role of index numbers in deriving theoretically meaningful measures of productivity. It is quite easy to see how the index number literature is closely connected with productivity measurement. For example, the Hicks-Moorsteen TFP index, defined as a ratio of the output and input quantity index numbers, can be made operational using the

[18] This does not imply or follow from the results concerning the translog form we obtained in Section 4.6.3 when dealing with output quantity index numbers.

results in this section. Similarly, if we wish to use profitability ratios and adjust them for price level differences, we need to make use of appropriate output and input price index numbers.

In the following section a simple numerical example with an artificial data set are computed and results are presented.

4.7 A Simple Numerical Example

We introduce a simple example which is used throughout the remainder of this chapter. Table 4.1 contains five years of price and quantity data from a hypothetical firm called Billy's Bus Company.[19] This company uses three inputs: labour, capital and "other", to produce two outputs: metropolitan passenger kilometres and long distance passenger kilometres.[20]

Table 4.1 Data for Billy's Bus Company

	INPUTS					
	quantity			price		
year	labour	capital	other	labour	capital	other
1990	145	67	39	39	100	100
1991	166	75	39	41	110	97
1992	162	78	43	42	114	103
1993	178	89	42	46	121	119
1994	177	93	51	46	142	122

	OUTPUTS			
	quantity		price	
year	metropolitan	long distance	metropolitan	long distance
1990	471	293	27	18
1991	472	290	28	17
1992	477	278	34	17
1993	533	277	32	20
1994	567	289	34	23

[19] It is noted that this example was inspired by the excellent Choochoo Railway example presented in Industry Commission (1992).

[20] Note that in most real applications we would have most probably begun with data on the values (i.e., expenditures or receipts) for all five variables, quantities for the two outputs and for the labour input, and price deflators (most likely obtained from a statistical agency) for the capital and "other" inputs. We would have then derived implicit quantities for the capital and "other" inputs and implicit prices for the two outputs and labour input.

The calculations in this example are conducted using the SHAZAM econometrics package (White, 1993), which has an INDEX command that calculates the Laspeyres, Paasche, Fisher and Törnqvist indices automatically. Note, however, that any standard spreadsheet program, such as EXCEL or LOTUS, or any statistical package, such as MINITAB or SAS, is also adequate for computational purposes. In fact, it is so easy to compute these indices it is best if a spreadsheet like EXCEL is used.

We illustrate the calculation of price and quantity indices by calculating the relevant output price and output quantity indices for this firm. The SHAZAM instructions are presented in Table 4.2a. The commands are quite simple. To calculate a price index, the INDEX command requires that the price and quantity data columns are listed together in pairs (i.e., price1, quantity1, price2, quantity2, etc.)

The SHAZAM output listing is presented in Table 4.2b. It is important to note that SHAZAM uses the term "Divisia" to describe the Törnqvist index.[21] This term is used because the Törnqvist index is a discrete approximation to the Divisia index. The output of the INDEX command lists the four price indices first, followed by four indirect quantity indices. Note that the price indices all have the value 1 in the first year (1990). This is the base year that is used by the INDEX command unless another base year is specified (for more on this, see White, 1993).

Table 4.2a SHAZAM Instructions for Output Price and Quantity Indices

	SHAZAM Command	**Description**
1.	sample 1 5	indicates that there are 5 observations
2	read yr q1 q2 p1 p2	read data on Year (Yr), 2 outputs quantities and prices
3 4 5 6 7	1990 471 293 27 18 1991 472 290 28 17 1992 477 278 34 17 1993 533 277 32 20 1994 567 289 34 23	data set - note that data can be read from a file instead of listing in the program.
8	** output price indices	comment line
9	index p1 q1 p2 q2	calculates chained price index numbers using different formulae
10	** output quantity indices	comment line
11	index q1 p1 q2 p2	calculates chained quantity index numbers using different formulae

[21] Note that the INDEX command in SHAZAM automatically produces a chained Divisia (Törnqvist) index. For all other formulae, we need to give the CHAIN option to create chained indices. Unless otherwise specified, indices for Laspeyres, Paasche and Fisher are calculated using the first period as the base period.

The INDEX command in SHAZAM can also be used to calculate direct quantity indices by simply changing the order that prices and quantities are listed on the command line. This is illustrated by the second INDEX command listed in Tables 4.2a and 4.2b. Be sure to note that SHAZAM incorrectly labels the quantity index as a price index and vice versa, in this instance.

Table 4.2b SHAZAM Output for Output Price and Quantity Indices

```
|_sample 1 5
 |_read yr y1 y2 p1 p2
   5 VARIABLES AND        5 OBSERVATIONS STARTING AT OBS       1

 |_** output price indices
 |_index p1 y1 p2 y2

 REQUIRED MEMORY IS PAR=    1 CURRENT PAR=  500
 BASE PERIOD IS OBSERVATION     1
                PRICE INDEX                          QUANTITY
DIVISIA  PAASCHE  LASPEYRES  FISHER   DIVISIA       PAASCHE    LASPEYRES
FISHER
1   1.000   1.000   1.000   1.000   0.1799E+05  0.1799E+05   0.1799E+05
0.1799E+05
2   1.010   1.010   1.010   1.010   0.1797E+05  0.1796E+05   0.1797E+05
0.1797E+05
3   1.169   1.171   1.167   1.169   0.1792E+05  0.1788E+05   0.1795E+05
0.1792E+05
4   1.159   1.166   1.163   1.165   0.1949E+05  0.1938E+05   0.1942E+05
0.1940E+05
5   1.256   1.264   1.265   1.264   0.2064E+05  0.2051E+05   0.2050E+05
0.2051E+05
 |_** output quantity indices
 |_index y1 p1 y2 p2

 REQUIRED MEMORY IS PAR=    1 CURRENT PAR=  500
 BASE PERIOD IS OBSERVATION     1
                     PRICE INDEX                          QUANTITY
     DIVISIA  PAASCHE  LASPEYRES  FISHER   DIVISIA      PAASCHE   LASPEYRES
FISHER
   1   1.000   1.000   1.000   1.000   0.1799E+05   0.1799E+05   0.1799E+05
0.1799E+05
   2   0.999   0.999   0.998   0.999   0.1817E+05   0.1817E+05   0.1817E+05
0.1817E+05
   3   0.996   0.998   0.994   0.996   0.2103E+05   0.2099E+05   0.2107E+05
0.2103E+05
   4   1.083   1.079   1.077   1.078   0.2086E+05   0.2093E+05   0.2098E+05
0.2096E+05
   5   1.147   1.139   1.140   1.140   0.2261E+05   0.2275E+05   0.2274E+05
0.2275E+05
```

4.8 Transitivity in Multilateral Comparisons

In this section, we consider the problem of deriving price and quantity index numbers over space at a given point of time. This problem arises when we are interested in output, input and productivity comparisons across a number of countries, regions, firms, plants, etc. In such cases, we are typically interested in all pairs of comparisons, i.e., comparisons across all pairs of firms. Suppose, we derive an index, I_{ij}, for a pair of firms (i,j), using a formula of our choice. We consider all pairs (i,j) with $i,j = 1,2,...,I$ where I represents the total number of firms. Then we have a matrix of comparisons between all pairs of firms,

$$\begin{bmatrix} I_{11} & I_{12} & \dots & I_{1I} \\ I_{21} & I_{22} & \dots & I_{2I} \\ \vdots & & & \\ I_{I1} & I_{I2} & \dots & I_{II} \end{bmatrix}. \tag{4.30}$$

This matrix represents all multilateral comparisons involving I firms and, ideally, we would like these comparisons to be internally consistent, i.e., to satisfy the property of transitivity.

Internal consistency requires that a direct comparison between any two firms i and j, should be the same as a possible indirect comparison between i and j through a third firm k. Thus, we require, for any i, j and k,

$$I_{ij} = I_{ik} \times I_{kj}. \tag{4.31}$$

For example, if a matrix of index numbers shows that firm i produces 10% more than firm k and firm k produces 20% more than firm j, then we should always find that firm i produces 32% (1.1×1.2=1.32) more than firm j.

Transitivity is an extremely important property to be satisfied by index numbers used in multilateral spatial comparisons of prices and quantities. When we make use of cross-sectional data for purposes of productivity comparisons, we need to compute input and output price and quantity index numbers for pairs of firms in the sample. Such pair-wise comparisons should be internally consistent and, therefore, need to satisfy transitivity. See Rao (2004) for an exposition of the issues involved in spatial comparisons.

Unfortunately, none of the index number formulae, including Fisher and Törnqvist, satisfy the transitivity property, given in equation (4.31). Remember, however, that these two indices do satisfy the time-reversal test: $I_{st} = 1/I_{ts}$.

The problem then is: how do we obtain consistent multilateral comparisons between firms? A simple solution is to generate transitive indices from a set of non-

transitive multilateral comparisons using a technique due to Elteto-Koves (1964) and Szulc (1964). This method is known as the EKS method.[22] Caves, Christensen and Diewert (1982*a*) uses the EKS method to derive multilateral Törnqvist indices that are transitive.

We illustrate the conversion of non-transitive indices into transitive indices as follows. Suppose, we start with Törnqvist indices, I_{ij}^{T}, for all pairs *i,j*. Then, for all firms i and j, we use the EKS method to convert the Törnqvist indices into multilateral Caves, Christensen and Diewert (CCD) indices by calculating:

$$I_{st}^{CCD} = \prod_{k=1}^{I} \left[I_{ik}^{T} \times I_{kj}^{T} \right]^{\frac{1}{I}} \tag{4.32}$$

These indices satisfy the following properties:

(i) I_{ij}^{CCD}, for $i,j = 1,2,...,I$, are transitive.

(ii) The new indices, I_{ij}^{CCD}, deviate the least from the original Törnqvist indices in a least-squares sense.

(iii) If we focus on quantity indices based on the Törnqvist formula then the CCD index in log-change form can be shown to be equal to

$$\begin{aligned} \ln Q_{ij}^{CCD} &= \frac{1}{I} \sum_{k=1}^{I} \left[\ln Q_{ik}^{T} + \ln Q_{kj}^{T} \right] \\ &= \frac{1}{2} \sum_{m=1}^{M} \left(r_{mi} + \bar{r}_m \right) \left(\ln q_{mj} - \overline{\ln q_m} \right) - \frac{1}{2} \sum_{m=1}^{M} \left(r_{mj} + \bar{r}_m \right) \left(\ln q_{mj} - \overline{\ln q_m} \right) \end{aligned} \tag{4.33}$$

where $\bar{r}_m = \frac{1}{I} \sum_{k=1}^{I} \omega_{mk}$ and $\overline{\ln q_m} = \frac{1}{I} \sum_{k=1}^{I} \ln q_{mk}$.

The formula in equation (4.33) is the form proposed in Caves, Christensen and Diewert (1982a) and this is the form used in most empirical analyses of total factor productivity measurement conducted during the past decade.

(iv) The formula in equation (4.33) has an intuitive interpretation. A comparison between two firms is obtained by first comparing each firm with the average firm and then comparing the differences in firm levels relative to the average firm.

[22] This method was suggested in the context of deriving transitive index numbers for comparisons of prices across countries. This procedure is currently one of the methods used by the OECD for generating internationally comparable statistics on gross domestic product and its components. Rao and Banerjee (1984). consider the EKS method in detail. Only an application of this method in the context of productivity comparisons due to Caves, Christensen and Diewert (1982a) is considered here.

Although the question of transitivity is quite important, the rationale behind the CCD multilateral index is rarely spelt out. In this section, we attempt to examine the main logic behind the CCD index. From this viewpoint, although equation (4.33) is the most popular form for a multilateral Törnqvist index, it is desirable to use the form in equation (4.32) as the root of the multilateral index.

Equation (4.32) provides an approach that can be applied to binary indices without detailed price and quantity data. To elaborate, suppose we have a matrix of binary Fisher or Törnqvist price and quantity indices. Then how do we generate transitive indices? The formula in equation (4.33) is not very useful in such cases because it requires all the basic price and quantity data. Even if basic data are not available, it is feasible to apply equation (4.33) and derive multilateral comparisons that are transitive.

It is not obvious from equation (4.32) how this procedure can be applied if the preferred index formula is different from the Törnqvist index. Suppose, we are working with Fisher index numbers for output index numbers between firms. Let Q_{ij}^{F} represent the Fisher index for *firm i* with *firm j* as base. Obviously, the Q_{ij}^{F}s for $i,j = 1,2,...,I$ do not satisfy transitivity. The EKS procedure in equation (4.32) can be applied to yield consistent indices as:

$$Q_{ij}^{F-EKS} = \prod_{k=1}^{I}\left[Q_{ik}^{F} \times Q_{kj}^{F}\right]^{\frac{1}{I}}.$$

The resulting quantity index numbers, $Q_{st}^{F\text{-}EKS}$, satisfy the transitivity property.

It is important to bear in mind that the condition of transitivity is an operational constraint preserving internal consistency. The imposition of the transitivity condition implies that a quantity (or price) comparison between two firms, *s* and *t*, is influenced by price and quantity data for not just these two firms but all the other firms in the analysis. Hence, the addition of an extra firm to the sample necessitates the recalculation of *all* indices. For a review of the methods used in constructing transitive multilateral comparisons across firms, see Rao (2004).

4.9 TFP Change Measurement Using Index Numbers

Above, we examine the mechanics of constructing price and quantity index numbers and in the theoretical underpinnings of some of the commonly-used index number formulae. In Section 3.5, we discuss the various approaches to the measurement of TFP change. To measure productivity changes, index numbers are used in measuring changes in the levels of outputs produced and the levels of inputs used in the production process over two time periods or across two firms. The focus of this section is to describe the computational methods used in deriving an index of TFP, either over time or across firms or enterprises. A TFP index may be applied to

binary comparisons, where we wish to compare two time periods or two cross-sectional units, or it may be applied to a multilateral situation where the TFP index is computed for several cross-sectional units.

4.9.1 Binary Comparisons

Consider first the Hicks-Moorstee TFP index for two time periods (or enterprises), s and t.[23][GEB6] Following Section 3.5, the HM TFP index in its logarithmic form is given by

$$\ln TFP_{st} = \ln \frac{\text{Output Index}_{st}}{\text{Input Index}_{st}} \tag{4.34}$$

In order to compute numerical values of this measure of TFP change, we need to use one of the index number formulae discussed in Section 4.3. Based on the test and economic-theoretic properties of index numbers, the most obvious choice is to use the Fisher index or the Törnqvist index to compute the input and output indices from the observed price and quantity data on outputs and inputs.

For the purposes of this section, we let $\mathbf{q}_s$ and $\mathbf{x}_s$ represent output and input quantities, and $\mathbf{r}_s$ and s_s represent the revenue shares of outputs and cost shares for inputs, respectively. Subscripts, s and t, stand for time periods (or firms), and m is used to denote the m-th output commodity and k to denote the k-th input commodity.

In most empirical applications, where TFP indices are calculated, the Törnqvist index formula is used for purposes of output and input index calculations. Then the Törnqvist TFP index[24] is defined, in its logarithmic form as

$$\ln TFP_{st} = \ln \frac{\text{Output Index}_{st}}{\text{Input Index}_{st}} = \ln \text{Output Index}_{st} - \ln \text{Input Index}_{st}$$

$$= \frac{1}{2}\sum_{m=1}^{M}(r_{is} + r_{it})(\ln q_{mt} - \ln q_{ms}) - \frac{1}{2}\sum_{n=1}^{N}(s_{ns} + s_{nt})(\ln x_{nt} - \ln x_{ns}), \tag{4.35}$$

where the first part of the right-hand side of equation (4.35) is the logarithmic form of the Törnqvist index applied to output data, and the second part is the input index, calculated using input quantities and the corresponding cost shares.

Equation (4.35) suggests that it is possible to replace the Törnqvist index by any other suitable formula. Diewert (1992) suggests the use of the Fisher index which

[23] In the case of cross-sectional data on firms, we use indexes *i and j* and the total number of firms is given by I.

[24] It appears that, in many cases, the Törnqvist TFP index is considered to be essentially the same as the TFP index. But we make a distinction between these two. Equation (4.27) defines a general TFP index and equation (4.28) is a particular case where indices are derived using the Törnqvist index formula.

has many desirable properties. In many respects, the Fisher index is more intuitive than the Törnqvist index and, more importantly, it decomposes the value index exactly into price and quantity components. The fact that it is in an additive format also makes the Fisher index more easily understood. In this case, the TFP index is given by

$$TFP_{st} = \frac{\text{Output Index}_{st}(\text{Fisher})}{\text{Input Index}_{st}(\text{Fisher})}.$$

Since the Fisher and Törnqvist index numbers both provide reasonable approximations to the true output and input quantity, indices in most practical applications involving time-series data, both formulae yield very similar numerical values for the TFP index (Diewert, 1992).

Malmquist TFP Index and Its Computation Using Törnqvist Index

So far, we have focused only on measures of TFP change using the Hicks-Moorsteen index. Here, we describe an important result, due to Caves, Christensen and Diewert (1982a, b), which shows that the Malmquist TFP index can be approximated, under a set of conditions, simply by the ratio of an output quantity index to an input quantity index, where both indices are computed using the Törnqvist formula. A similar result relating to the Fisher index was provided in Diewert (1992).

We state, without proof, the following result from Caves, Christensen and Diewert (1982a).

Result 4.9: If the Malmquist output distance functions for periods s and t have a translog functional form with identical second-order terms, then, under the assumption of technical and allocative efficiency of the firm in the two periods, the geometric average of the two output-based Malmquist TFP productivity indices in equations (3.45) and (3.46) is given by

$$m_o(\mathbf{q}_s,\mathbf{q}_t,\mathbf{x}_s,\mathbf{x}_t) = \left[m_o^t(\mathbf{q}_s,\mathbf{q}_t,\mathbf{x}_s,x_t)\times m_o^s(\mathbf{q}_s,\mathbf{q}_t,\mathbf{x}_s,\mathbf{x}_t)\right]^{0.5}$$

$$= \frac{\text{Tornqvist output index}}{\text{Tornqvist input index}} \times \prod_{n=1}^{N}\left(\frac{x_{nt}}{x_{ns}}\right)^{s_n^{*}/2} \tag{4.36}$$

where $s_n^* = s_{nt}(1-\varepsilon_t) + s_{ns}(1-\varepsilon_s)$; ε_t and ε_s are the local returns-to-scale values in periods t and s, at the observed input and output levels, respectively; and s_n^s and s_n^t are the shares of n-th input in total input cost in periods s and t, respectively.

Algebraic formulae for the Törnqvist output and input indices are given Section 4.3. These indices can be computed using the observed input and output quantities and shares in the two periods.

Result 4.9 is an extremely useful result. A few remarks are in order.

- First, it shows that, even when we do not know the exact nature of the output distance functions, we can define an exact measure of the geometric average of the Malmquist output-orientated productivity indices, based on the technologies of periods s and t.

- If we have constant returns to scale in both periods ($\varepsilon_t = \varepsilon_s = 1$), then

$$\text{Malmquist TFP index} = \frac{\text{Tornqvist output index}}{\text{Tornqvist input index}}.$$

  This is the total factor productivity measure used in most empirical studies, and it coincides with the Hicks-Moorsteen TFP index.

- Under decreasing returns to scale, using duality results and a profit-maximisation assumption, Caves, Christensen and Diewert (1982b) show that the returns-to-scale parameters can be measured using the observed price and quantity data as:

$$\varepsilon_j = \frac{\text{value of output in period } j}{\text{value of input in period } j}, \quad j = t, s.$$

  In the case of increasing returns to scale, observed costs and revenues cannot be used to compute returns–to-scale parameters.

- Equation (4.36) shows that the second factor on the right-hand side depends, not only upon the local returns-to-scale parameters, ε_t and ε_s, but also upon the relative input levels in periods t and s. If the input use has not changed over the two periods, then returns-to-scale issues do not arise in productivity change calculations.
- Result 4.9 provides an economic-theoretic justification for the use of the standard measure of total factor productivity (TFP), defined as a ratio of Törnqvist indices of output and inputs. Such a justification holds when the underlying technologies exhibit constant returns to scale.

We conclude this discussion by making several points. First, it is easy to see that the CCD result can also be applied to the input-orientated Malmquist TFP index with suitable modifications to equation (4.36). Second, the CCD results suggest that, if we are able to make some assumptions on the behaviour of firms, then it is not necessary to actually calculate all the output and input distances involved in the

definition of the Malmquist TFP index. Third, the CCD assumptions also suggest a natural interpretation of the Malmquist TFP index. Under their assumptions, CCD essentially measures technical change as there are no technical or allocative efficiency assumptions and scale of operations is not an issue because of the assumption of constant returns to scale. Finally, we point out that index numbers have an important role in measuring TFP change over time and TFP levels across firms. With limited data, we are able to obtain measures of TFP. However, the CCD results also imply that where our data suggests that either the prices are not market prices or the behaviour of the firms may not be optimal, then use of the index number approach to TFP productivity measurement may not be measuring the Malmquist TFP index and no real economic-theoritic interpretation can be accorded to the input and output quantity index numbers.

4.9.2 Multilateral Productivity Comparisons

In the case of productivity comparisons across a number of firms, it is necessary to impose the transitivity condition on the index numbers used. In such cases, we recommend that the TFP indices in the previous section are computed using transitive EKS-type index numbers for measuring differences in the levels of inputs and outputs across firms. Rao (2004) provides a summary of the index number issues involved in deriving multilateral index numbers.

If the Hicks-Moorsteen approach is used, it is necessary to first generate transitive output and input quantity index numbers, based on the Fisher or Törnqvist indices, and then a ratio of the transitive indices be used to measure TFP change. A similar approach needs to be adopted in the case of the Malmquist TFP index.

Routine application of the formulae in equation (4.35) to multilateral comparisons involving more than two enterprises leads to the problem of transitivity. Application of a binary TFP index formula yields inconsistent results. Following Caves, Christensen and Diewert (1982*a*), and the discussion in Section 4.8, the following formula is used if the Törnqvist index formula is preferred. The following index is derived by applying the EKS approach to obtain a transitive CCD index, which is a multilateral generalisation of the Törnqvist index.

$$\begin{aligned}
&\ell n\ \mathrm{TFP}_{st}^{*} \\
&= \left[\frac{1}{2}\sum_{m=1}^{M}\left(r_{mt}+\bar{r}_m\right)\left(\ln q_{mt}-\overline{\ln q_m}\right)-\frac{1}{2}\sum_{m=1}^{M}\left(r_{ms}+\overline{r_m}\right)\left(\ln y_{ms}-\overline{\ln y_m}\right)\right] \\
&\quad -\left[\frac{1}{2}\sum_{n=1}^{N}\left(s_{nt}+\bar{s}_n\right)\left(\ln x_{jt}-\overline{\ln x_{nk}}\right)-\frac{1}{2}\sum_{k=1}^{K}\left(s_{ns}+\bar{s}_n\right)\left(\ln x_{ks}-\overline{\ln x_{ni}}\right)\right], \qquad (4.37)
\end{aligned}$$

where TFP_{st}^{*} is a transitive TFP index;

$\bar{r}_m$ is the arithmetic mean of the output shares;

$\bar{s}_n$ is the arithmetic mean of the input shares;

$\overline{ln\mathrm{y}_{\mathrm{m}}} = \frac{1}{\mathrm{I}}\sum_{\mathrm{i}=1}^{\mathrm{I}} ln\mathrm{y}_{\mathrm{mi}}$; and

$\overline{\ln x_n} = \frac{I}{I}\sum_{i=1}^{I} \ln x_{ni}$.

All averages are taken over I enterprises or time periods or a combination of both.

The formula in equation (4.37) is computationally simple and employed in many empirical TFP studies. However, following the more general definition of a TFP index in equation (4.37), we can define alternative TFP formulae by using transitive output and input indices in the general multilateral TFP index given by

$$TFP_{st}^{*} = \frac{\text{Transitive output index}}{\text{Transitive input index}} . \tag{4.38}$$

It is feasible to use any output and input index numbers of our choice in equation (4.38). A suitable choice is the multilateral generalisation of the Fisher index derived using the EKS procedure, discussed in Section 4.8.

In concluding this discussion of multilateral TFP measurement, we note that transitivity is an important requirement for spatial comparisons, but it is not a major problem in the case of temporal comparisons, where the observations are in a naturally ordered sequence in which a simple chain-base index would be adequate.

4.9.3 A Simple Numerical Illustration: TFP Computations

Continuing on with our Bus Company example, we can easily obtain a Törnqvist TFP index by finding the ratio of the output quantity and input quantity indices. These calculations could be performed using SHAZAM, as before, or one could use spreadsheet software, such as EXCEL or LOTUS. However, we use TFPIP Version 1.0, which is a computer program recently developed for the purpose of computing index numbers for input and output quantities, as well as the resulting TFP index numbers. The TFPIP program produces index numbers calculated using either the Fisher or the Törnqvist formulae.[25] The program produces chained index numbers for comparisons over time. An option to produce transitive multilateral comparisons is also available in TFPIP Version 1.0. The TFPIP program is discussed in the Appendix.

To calculate the TFP index for the data on Billy's Bus Company, we need to prepare the data and the instruction files (which must be text files) for using TFPIP. The data file for this example is given in Table 4.3a below. The data file consists of five annual observations, each row representing a year's observation. The first two columns represent quantities of the two outputs in the example, the next three columns show the input quantities. Price data are in the last five columns, again the first two columns representing the two output commodities and the last three columns containing the prices of the input commodities.

Table 4.3a Listing of Data file, EX1.DTA

471	293	145	67	39	27	18	39	100	100
472	290	166	75	39	28	17	41	110	97
477	278	162	78	43	34	17	42	114	103
533	277	178	89	42	32	20	46	121	119
567	289	177	93	51	34	23	46	142	122

The instruction file is listed in Table 4.3b below. Comments provided on each line explain the meaning of each instruction. The first two lines identify the location of the data and output files. The last two rows allow us to pick the formula we wish to use and the type of comparisons we wish to make.

[25] It is anticipated that future versions of this program will have the capability of producing index numbers derived using different formulae.

Table 4.3b Listing of Instruction File, EX1.INS

```
ex1.dta        DATA FILE NAME
ex1.out        OUTPUT FILE NAME
5              NUMBER OF OBSERVATIONS
2              NUMBER OF OUTPUTS
3              NUMBER OF INPUTS
0              0=TÖRNQVIST AND 1=FISHER
0              0=NON-TRANSITIVE AND 1=TRANSITIVE
```

Table 4.3c Listing of Output File, EX1.OUT

```
Results from TFPIP Version 1.0

Instruction file = ex1.ins
Data file        = ex1.dta

Törnqvist Index Numbers

These Indices are NOT Transitive


INDICES OF CHANGES REL. TO PREVIOUS OBSERVATION:

 obsn    output     input       TFP
    2    0.9986    1.1007    0.9073
    3    0.9974    1.0297    0.9686
    4    1.0877    1.0896    0.9983
    5    1.0586    1.0627    0.9962


CUMULATIVE INDICES:

 obsn    output     input       TFP
    1    1.0000    1.0000    1.0000
    2    0.9986    1.1007    0.9073
    3    0.9960    1.1333    0.8788
    4    1.0833    1.2348    0.8773
    5    1.1468    1.3122    0.8740
```

The output generated from the execution of TFPIP Version 1.0 is given in Table 4.3c. The output from the program lists output and input quantity index numbers as well as the resulting TFP index for each year, computed using the previous year as the base. These annual change index numbers are linked to provide chained (or cumulative) output, input and TFP indices with period 1 as the base period. Observe that the TFP index for this bus company has declined over the study period by almost 13%.

Computation of Transitive TFP Indices

We now return to the Billy's Bus Company example . The data set used in this example is the same as that in Table 4.3a above. The multilateral TFP index in equation (4.29), based on the Törnqvist index, is computed using TFPIP Version 1.0. The Instruction file for this computation is given below in Table 4.4a.

Table 4.4a Listing of Instruction File, EX2.INS

```
ex1.dta       DATA FILE NAME
ex2.out       OUTPUT FILE NAME
5             NUMBER OF OBSERVATIONS
2             NUMBER OF OUTPUTS
3             NUMBER OF INPUTS
0             0=TÖRNQVIST AND 1=FISHER
1             0=NON-TRANSITIVE AND 1=TRANSITIVE
```

Note from the listing above that the last instruction provides the option to derive a multilateral TFP index that is transitive. The second last line allows the choice between Törnqvist and Fisher indices.

Table 4.5b lists the output derived using the instruction file above.

Table 4.4b Listing of Output File, EX2.OUT

```
Results from TFPIP Version 1.0

Instruction file = ex2.ins
Data file        = ex1.dta

Törnqvist Index Numbers

These Indices are Transitive


INDICES RELATIVE TO FIRST OBSERVATION:

 obsn     output      input        TFP
    1     1.0000     1.0000     1.0000
    2     0.9979     1.1003     0.9069
    3     0.9938     1.1333     0.8769
    4     1.0792     1.2347     0.8741
    5     1.1417     1.3129     0.8696
```

Since all the index numbers are transitive, only one set of index numbers with the base period equal to 1 is presented. If, however, we wish to compute an index for period 4 with period 3 as base, for example, we simply divide the index in the table above for year 4, 1.0792, by the index for period 3, 0.9938. The table shows a steady decline in the TFP over the five-year period.

4.10 Empirical Application: Australian National Railways

We now present a real example based on a study on the Australian National Railways undertaken by the Industry Commission (IC). The example is drawn from Industry Commission (1992), which reports a number of studies measuring the TFP of government trading agencies in Australia.[26] The IC report describes a number of case studies on various enterprises including Australian National Railways, State Rail Authority of New South Wales, Melbourne Water, Port of Brisbane Authority, Pacific Power and Australia Post. We select the study on Australian National Railways for detailed discussion.

This is a study of TFP changes, at an aggregate level, over the period 1979/80 to 1990/91.[27] The main purpose of the study was to go beyond the usual partial productivity measures and construct a TFP index as a part of a set of key performance indicators.

The study covers a wide range of outputs and inputs. Three categories of outputs are distinguished, namely: mainland freight services, measured in net-tonne-kilometres (NTKs); Tasrail freight services,[28] measured in net-tonne-kilometres (NTKs); and passenger services, measured in passenger-train-kilometres (PTKs). The aggregated output data for the three categories along with the price data are presented in Table 4.5. These quantities are, in themselves, aggregates of distinctly different categories of output. For example, passenger output in kilometres ignores the class of travel as well as the terminal services provided. But the prices are also average prices, averaged over passenger travel of different kinds.

The study distinguishes between two types of inputs in computing the aggregate index, namely, capital inputs and non-capital inputs. Capital inputs are further divided into: land, buildings/structures and perway; plant and equipment; and the rolling stock. The non-capital inputs are divided into: labour; fuel; and "other inputs", a composite category.

---

[26] The authors are grateful to the Industry Commission for publishing this report without copyright. The example presented here is a summary version of a more detailed presentation in Chapter 4 of Industry Commission (1992).

[27] Years here are of the form 1979/80 etc., reflecting the accounting year used in Australia that extends from 1 July to 30 June.

[28] Tasrail refers to railway operations in the island state of Tasmania.

Table 4.5 Output Data for the Australian National Railways Example

Quantities			Prices		
Mainland Freight (1,000 NTKs)	**Tasrail Freight (1,000 NTKs)**	**Passenger (1,000 PTKs)**	**Mainland Freight ($/NTK)**	**Tasrail Freight ($/NTK)**	**Passenger ($/PTK)**
5235000	383000	2924	0.02	0.07	10
5331000	420000	3057	0.03	0.07	12
5356000	375000	2992	0.03	0.08	14
4967000	381000	2395	0.03	0.08	18
5511000	401000	2355	0.03	0.08	20
5867000	403000	2188	0.03	0.08	22
6679000	402000	2486	0.03	0.09	23
6445000	429000	2381	0.03	0.09	23
7192000	455000	2439	0.03	0.09	23
7618000	459000	2397	0.03	0.08	26
7699000	413000	2316	0.03	0.11	32
7420000	369000	1664	0.03	0.12	47

The labour input for the study is defined to be the amount of labour used during the year for operational and maintenance purposes. The quantity measure of labour refers to the level of full-time staff, as on 30 June. The study points out the need to refine the measure to capture the number of hours worked, as well as to make adjustments for the labour input used in producing capital stock. The labour input figures are shown in Table 4.6.

The second non-capital input used is the quantity of fuel used. The last of the inputs refers to the "other inputs" category, which is in the form of a series of *real expenditures* on this category. The price used for this item is the implicit price deflator for non-farm gross domestic product.

Table 4.6 Non-capital Input Data for the Australian National Railways Example

Quantities			Prices		
Labour (persons)	Fuel (1,000 litres)	Other Inputs ($1,000)[a]	Labour ($/person)	Fuel ($/litre)	Other Inputs (index)
10481	77380	119113	13097	0.18	0.45
10071	80148	112939	14730	0.26	0.50
9941	77105	108263	16692	0.28	0.56
9575	72129	110210	18651	0.37	0.62
9252	85868	109292	20166	0.37	0.66
8799	89706	97594	21307	0.39	0.70
8127	96312	93178	24990	0.41	0.75
7838	92519	80054	26412	0.42	0.81
7198	96435	77716	28572	0.43	0.87
6648	101327	74147	32617	0.39	0.94
6432	98874	80826	34565	0.43	1.00
5965	96016	73172	35646	0.46	1.04

Note: a. These quantities are in 1989/90 values.

Data for the non-capital inputs are drawn mainly from annual reports. These reports also form the basis for the output series discussed earlier. Table 4.6 shows the input quantity series as well as the prices. A major feature of this input series is the significant reduction in labour input over the study period. Both labour and "other inputs" recorded a decrease while fuel recorded a steady increase until 1988/89.

The study explicitly recognises the importance of measuring the capital input properly. Though the most appropriate input measure is the flow of capital services per period, a measure of capital stock is used in its place. The capital stock series for the three components were constructed using the Perpetual Inventory Method. The capital stock input used in the study refers to Australian National's own capital stock. Capital items leased were accounted for in the "other inputs" category. The price used for the capital items represents the economic rental price of capital. Capital inputs and the associated "prices" are shown in Table 4.7.

Table 4.7 Capital Input Data for the Australian National Railways Example

Quantities			Prices		
Land, Building and Perway ($1,000)[a]	**Plant and Equipment ($1,000)[a]**	**Rolling Stock ($1,000)[a]**	**Land, Building and Perway (index)[b]**	**Plant and Equipment (index)[b]**	**Rolling Stock (index)[b]**
1858038	94057	332307	10	50	50
2101035	93927	308491	20	80	80
2059365	89764	285626	30	120	120
2118357	93271	269265	30	100	100
2117625	91837	275134	70	140	140
2095680	90120	261495	70	160	160
2069494	89617	251588	50	90	90
2034867	88773	239736	70	120	120
2017626	89653	235834	80	200	200
1998345	98762	252514	80	240	240
2011753	100495	251850	80	190	190
2018802	107654	242662	130	200	200

Notes: a. These quantities are in 1989/90 values.
b. These are indices of the rental price of capital.

Table 4.8 shows the aggregate output and input indices as well as the total factor productivity index derived using the TFPIP Version 1.0 program. The aggregate inputs index for the period has shown a steady decline, by about 25 per cent. A significant factor to consider is that the input decline occurred at a time when output has shown an increasing trend.

The output data are combined using the Törnqvist index formula, to compute the output index presented in Table 4.8. These indices are also presented in Figure 4.4. Over the study period, the output of Australian National has shown a steady increase, but with downturns in 1982/83 and also towards the end of the study

period. The first of these downturns was associated with severe drought in rural Australia and the latter downturn shows the effects of the recession in the economy.

Table 4.8 Indices of Output, Input and TFP for Australian National Railways

Year	Output	Input	TFP
79/80	1.0000	1.0000	1.0000
80/81	1.0343	0.9782	1.0573
81/82	1.0188	0.9515	1.0707
82/83	0.9304	0.9345	0.9956
83/84	1.0014	0.9316	1.0748
84/85	1.0311	0.8950	1.1521
85/86	1.1543	0.8596	1.3428
86/87	1.1268	0.8191	1.3756
87/88	1.2293	0.7885	1.5590
88/89	1.2766	0.7690	1.6600
89/90	1.2607	0.7684	1.6407
90/91	1.1283	0.7376	1.5296

Combining the input and output indices to derive the TFP index, we see that, apart from a decline during 1982/83 and then over the most recent period associated with recession in the Australian economy, TFP showed a compound rate of growth of four per cent. This growth was significantly above the growth rate experienced in the whole economy.

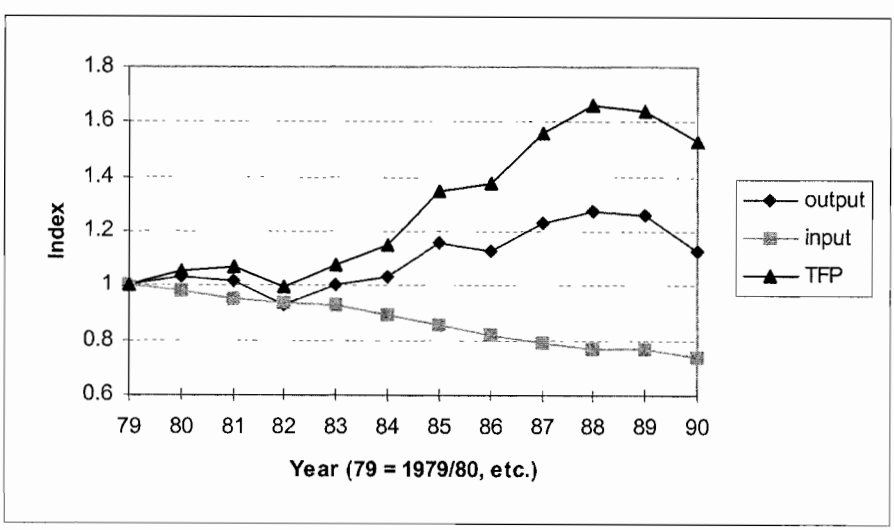

Figure 4.4 Indices of Output, Input and TFP for Australian National Railways

Industry Commission (1992) provides some discussion of the TFP growth measures obtained for the Australian National Railways. This growth can be attributed to both *efficiency gains,* derived through labour reductions and internal reorganisation, as well as the effects of *technical progress.*

Given the aggregated nature of the data, it is not possible to measure these two components. The study speculated that efficiency gains were the main reason for the high compound growth rate of the TFP index. The Australian National Railways TFP growth study illustrates the various steps involved in the compilation of the TFP indices and also its shortcomings when it comes to final interpretation and use.

4.11 Conclusions

In this chapter, we deal with various index number formulae that are generally used in the construction of price and quantity index numbers. However, the main purpose of the chapter is to explain how index number methods play a significant role in computing various TFP change indices that are discussed in Chapter 3. Thus, this chapter demonstrates how, using very limited data, we can compute various measures of TFP change and the underlying assumptions that make it possible. We emphasise that proper care should be taken in interpreting the resulting TFP measures. If we invoke the result of CCD, then the index-number based TFP measures technical change. However, if these assumptions do not hold, we can still compute the TFP index as a measure of TFP change but the interpretation of the index is not straightforward. In such cases, we need to find a way of computing all the distances involved in defining either the Malmquist TFP index or the source or component-based measure of TFP change, proposed in Balk (2001). In both of these cases, we need a lot more data than just two observations, one for each time period, to be able to measure TFP change. We need data on a reasonable number of firms or cross-sections for the two periods under consideration, to make it possible to use the techniques such as the data envelopment analysis (DEA) or the stochastic frontier analysis (SFA) methods. These techniques are the subject matter of Chapters 6 to 10. In Chapter 12, we provide a comparison of various methods for measuring TFP change and we compare and contrast the index number approach with the DEA and the SFA methods.

A final comment before we end this chapter is in order. Throughout this chapter we have focused only on multiplicative index numbers, mainly due to the fact that all the efficiency and productivity measures are all multiplicative. Much of the traditional index number theory and practice concerns multiplicative index numbers. However, in the recent years there have been some applications of additive index numbers. Diewert (1998) describes additive index numbers which are defined in the form of differences rather than ratios, as in the case of multiplicative index numbers, and draws attention to the Bennet (1920) index which is an additive index number. Diewert (2000) shows how productivity measurement using differences can be undertaken. The theoretical framework underlying additive index numbers is based

on directional distance functions and these are too advanced for detailed discussion in this book. We refer interested reader to the work by Diewert (1998, 2000) and the recent monograph by Färe and Grosskopf (2004). Additive index numbers have very interesting applications to the economic approach to profit change decomposition.

5. DATA AND MEASUREMENT ISSUES

5.1 Introduction

The main objective of this chapter is to discuss the principal issues involved in data compilation with a special focus on the choice of variables, construction of variables for analysis and identifying the correct sources of data. In Chapter 4 we described index number methods used in compiling price and quantity index numbers that can be used in productivity measurement. The index number methods discussed earlier also have a major role to play in data compilation and management. Chapters 6 to 10 are devoted to data envelopment analysis (DEA) and stochastic frontier analysis (SFA) methods for the measurement of productivity and efficiency. These techniques should be used in conjunction with carefully compiled data on input and output quantities and prices.

We stress that quality and appropriateness of data used in these sophisticated techniques are just as important as the techniques themselves. There is much truth and significance in the old adage in empirical economics, "garbage in equals garbage out". It does not matter how powerful a given statistical technique or mathematical tool may be, it cannot overcome problems that fundamentally reside in the data themselves. Which variables to use? Are the selected variables consistent with the concept and phenomena they are supposed to capture? What sources should be used for which variables? What is the reliability of data that are compiled? Are there measurement errors and outliers? These are questions that an applied economist or econometrician embarking on a productivity measurement exercise should always ask themselves. A comprehensive assessment of the alternatives and data sources and a preliminary data editing exercise should precede any full-scale empirical analysis involving DEA or econometric estimation SFA models.

In the context of efficiency and productivity measurement, three categories of variables are important. These are: (i) output quantities; (ii) input quantities; (iii) prices of outputs and inputs; and (iv) quality characteristics of the various input and output variables are also important. From Chapters 2 and 3, it is clear that productivity measurement is based on a specific set of inputs used in the production of a specific set of outputs. Production technology is viewed as the transformation of a vector of inputs into a vector of outputs. Prices play a role in the determination of the composition of the outputs and inputs that maximise revenues or costs or profits and, therefore, are quite important in estimating dual econometric models. Price data are also central to the idea of allocative efficiency. In addition, price data play a significant role in aggregating data and in deflating value aggregates to derive constant price or real aggregates.

The types of problems encountered in the process of compiling appropriate data vary with each empirical exercise. These vary according to whether the study involves a micro- or macro–level analysis. These problems also depend upon the particular sector involved: agriculture; manufacturing; or the service sector. Further, problems in identifying the right output indicators and/or output measures may be a problem when one is dealing with the service sector. Different problems are encountered when dealing with firms operating in the market and non-market sectors of the economy. In the case of the market sector, it may be a lot easier to identify appropriate output price measures whereas in the non-market sector output price data may not exist.

We are conscious of the difficult nature of writing a prescriptive chapter which provides clear guidelines that are appropriate for all situations. Keeping this in perspective, we organise material in this chapter that allows us to highlight the principal issues in a systematic way. First, we discuss various issues relating to efficiency and productivity measurement involving a cross-section of firms or decision making units. In the material presented below, we make use of two examples to illustrate the points raised. We consider firms or factories producing television sets as the first example and universities offering tertiary education as the second example. The first is a standard firm producing goods that are sold in the market whereas the second example relates to educational institutions providing services to students and the society that are not provided to the uses for a stipulated price. We then proceed to briefly discuss data for productivity comparisons over time, at the sectoral level and for comparisons across countries and regions.

This chapter draws heavily from the material contained in OECD (2001a), which is the OECD Manual on measuring productivity; Morrison-Paul (1999) and Coelli *et al.* (2003).

The outline of the chapter is as follows: Sections 5.2 and 5.3 deal with input and output variables that are representative of the production activities of firms and enterprises; problems associated with prices are considered in Section 5.4; and the

last section is devoted to a discussion of some very basic data editing procedures recommended for use prior to the proper analysis.

5.2 Outputs

Measurement of outputs is probably one of the easiest tasks when the firms under consideration are commercial entities that produce tangible goods and services that are sold in the market place. It is more difficult to identify the outputs of an enterprise involved in delivering services, in particular, in the non-market sector. For example, while it is fairly straightforward to identify and then measure the outputs of a firm producing television sets, it is more difficult to do the same in the case of universities. It is much easier to identify and measure the output of a firm than to obtain suitable measures of output of a sector or a country.

Output measures for firm-level analysis

Measuring outputs at a firm level is more straightforward. In general, we have two possibilities, firms producing a single product and those producing multiple products. These are considered below.

Single-output firms

The outputs of single-product firms are the simplest to handle. In such a case, output for a firm is measured by the number of units produced in a calendar year. Often data on value of production or sales of the firms are available. These can be used as measures of output. However, even in such a simple case, a few problems could arise that need proper consideration.

- In many cases, published output data for firms may be given in terms of sales during the year. In this instance, we need to estimate the actual production during the year by adjusting the sales data by any change in inventories that may have occurred during the year (i.e., changes in the stock balance between the beginning and the end of the year). If the
- A more difficult issue concerns the quality differences embodied in the goods produced by different firms included in the study. Suppose we are considering the productivity of firms manufacturing TV sets, then it is not enough if output is measured by the number of TV sets produced by different firms. As TV sets vary in their specifications quality characteristics, it is necessary to measure output by different types of TV sets or to make some quality adjustment for differences in the overall quality of the product across different firms.
- In many practical applications, sales of different firms are provided in the total value of sales recorded in a year. After adjusting the sales for changes in inventories, we can get an estimate of the total value of production. If all the firms operate in a market of a single price then the nominal values can

be considered as an adequate measure of output. In the case where firms face different prices, it is essential that the value figure is converted using the price information. If only one good with uniform quality is involved, this is not a major issue.

Multiple-output firms

The case of firms producing multiple outputs can be a little more complex. The simplest case is where output quantities are available for all the products produced by the firms. If the number of products involved is not large, then it is a simple matter to use the output data in deriving measures of productivity and efficiency, subject to a satisfactory resolution of some of the issues raised in the case of single-output firms. Often the main concern here is if the product list is too long, in which case it may create the problems of degrees of freedom – similar to that encountered in standard regression analysis. To illustrate this, suppose we have output and input data collected from 30 firms producing TV sets for purposes of productivity comparisons. Suppose these firms produce and sell 50 different types of TVs identified by differences in the screen size, whether they are normal or plasma or LCD screens and if the TV sets are digital. If we treat these as 50 different products, then it is difficult to estimate the underlying frontiers with only 30 observations. In such cases, it is necessary to aggregate the data to form a smaller number of output aggregates – usually into two or three categories. The problem then is how to aggregate detailed quantity information? Index number methods are often used. Some important issues to consider in this situation are listed below.

1. As the aggregates are used in measuring productivity across firms or over time, it is important to ensure that the aggregates formed are meaningful. It is important that we do not aggregate oranges and apples but may aggregate over different varieties of apples or oranges. There are some guidelines to check when it is appropriate and admissible to aggregate over commodites. These are known as the Hicks and Leontief conditions for aggregation. Usually, it is necessary to ensure that the aggregates are formed across products that exhibit similar movements in relative prices or quantities or they meet separability conditions with respect to the production function. which requires that the relationship between two commodities, measured by their ratio of marginal products, does not change with a change in relative prices. The main message here is that care must be exercised in aggregating outputs of very different types of commodities.

2. We note that price data are needed in deriving these aggregates. Therefore, price data are an integral part of the work when multiple outputs are considered. Once we are satisfied that it is appropriate to form an aggregate of outputs of a number of products, then it is a simple matter to form a value aggregate by taking the product of price and quantity and summing it over all the commodities included in the aggregate.

3. If value aggregates, such as total value of TV sets produced by firms or total revenue generated by the firm, are being used as measures of output for a composite commodity, then it is necessary to make an adjustment for differences in the levels of prices. This means that it is necessary to deflate the nominal value aggregate using an appropriate price index. However when considering a large number of firms which are located in different regions within a country, it may be necessary to construct a *transitive* price index using the EKS formula discussed in Chapter 4.

4. If we have all the necessary data on prices and quantities it is easy to construct real value aggregates through the use of a properly constructed price index. Such detailed data may not always be available. In such instances, we need to pick an appropriate index as a proxy. In the case of TV industry, a price index for TVs or electronics goods may be used as a substitute for the price index.

Output measures for service industries

Performance measurement of service providers, both for profit and not-for-profit enterprises, is of considerable interest. Benchmarking the performance of hospitals, aged-care facilities, schools, banks, railways, airlines, police stations, utilities and other industries has attracted considerable interest over the last three decades. This is attributable to the availability of methods such as the data envelopment analysis (DEA) that can be applied where the output measures are in the form of output indicators, and in cases where price data are either not available or not relevant, as is the case for many non-marketed services.

Let us consider the case of universities. What are the output measures that accurately reflect the role and function of a university? We may consider teaching, research and community service as the three main activities of a university. We need to identify output measures associated with each of these activities. In terms of teaching we may consider the following output measures:

- Number of students
- Number of full-time equivalent students – this measure accounts for part-time and full-time students
- Number of students in different disciplines – in social sciences, sciences, medicine and engineering
- Number of undergraduate and postgraduate students
- Number of *weighted* full-time equivalent students – it weights students by the discipline and at the level

If all the universities under consideration have a similar student-mix then the choice of the output measure may not make too much difference to the productivity measure.

Now let us consider the second activity – research. Research output of a university may be measured using different variables including:

- Number of research publications – number of books, book chapters, journal articles, conference papers and other scholarly publications;
- Number of weighted research publications – each publication is weighted depending upon whether it is a book or a chapter or a journal article – it is necessary to specify some weights in constructing this measure;
- Quality adjusted research publications – this aggregate is formed by weighted each publication by the quality of the journal in which it is published;
- Total research grant income received – though this variable is considered as a research output there could be an argument as to whether grants are an input into research or an indicator of research output; and
- Total number of postgraduate students trained during a given year.

It is necessary to select one or two of these indicators to measure research output. We may observe here that there may be a strong correlation between these variables in which case choice of one variable over another may not have serious consequences. In the case of universities, it may be more difficult to identify suitable measures of community service and involvement. Some of the common measures include the number of radio interviews or expert opinions given by staff members of a university.

Just to give an indication of the type of variables researchers use in assessing the performance of various service industries, the following is a select list of variables seen in various empirical studies: number of patient days in hospital; number of different types of procedures performed; number of resident days in an aged-care facility; number of low-care and high-care residents in an aged-care facility; number of students in a school during a year; number of full-time equivalent students at a university in a given year; research output of a university, as measured by the number of research publications and the amount research grants received; cash deposits and total lending at different branches of a particular bank; number of accounts services at the branches; number of passengers carried by railways; total freight transported by railways; number of passengers carried by an airline; population of towns serviced by a police station; number of foot-patrols conducted; number of passenger kilometres travelled; and the quantity of electricity supplied. Various other indicators can be seen in published research on efficiency and productivity of various service enterprises.

While it difficult to deal with all the issues relating to the choice of indicators, a few important points need to be considered in choosing suitable output measures.

- Selection of appropriate output measures that accurately reflect the output of a particular enterprise is very important. For example, a simple measure like the total number of students studying at a university may not accurately

reflect the teaching activity of a university. Similarly, a simple measure like the number of passengers or the total volume of freight transported by trains may not be appropriate as the distances associated with such transportation is very important.

- Determining the appropriate level of aggregation and choice of weights for such aggregation are quite important. For example, one may consider the total number of students (full-time equivalents) as one of the measures of output of a university. However, students may be training in different areas, such as the social sciences, sciences, engineering and technology and medicine. Resource requirements to provide training in different areas require vastly different resource levels. In such cases, it may be necessary to either use disaggregated data or use aggregated data where appropriate weights are used in aggregating students in different disciplines.

- Ordinal measures, such as a ranking index in the range of 0 to 100, where the values of the index provide an indication of only ranking rather than the differences in actual levels, cannot be used as output measures. It is too easy to simply use such numbers and forget that any other rank-preserving transformation of an ordinal measure is equally acceptable, and there is no way of choosing between these two, and, more importantly, these two measures result in different measures of efficiency and productivity. Here the important consideration is the cardinality versus ordinality properties of the output measure used. Given the production theory basis, it is understood that output measures are cardinal measures so that the levels and differences are important and meaningful.

- It is also important to ensure that the choice of the output variables is consistent with the axioms of production technology discussed in Chapters 2 and 3. In particular, it is necessary to pay particular attention to the axioms of monotonicity and the convexity properties of the output sets.

- The final point concerns the treatment of quality. Most of the output measures listed above are broadly specified and in some cases quality of the output of a particular enterprise can differ greatly relative to another enterprise. The quality issue is discussed below in greater detail.

Output measures and quality differences

The discussion thus far has not explicitly referred to the problem of quality associated with different output indicators listed above. For example, two firms producing TV sets may be producing sets with completely different quality characteristics. Similarly, It is possible that two different universities may have trained the same number of students during a given year but the quality of outcomes associated with the skills attained and the level of satisfaction derived by the students may differ.

Quality is a multi-dimensional phenomenon that is associated with a commodity or service produced by a given enterprise. In addition to the common characteristics associated with the quality of a good, such as the class of travel (economy or first) or the length of travel associated with a journey, one may wish to consider other quality aspects of railway travel such as punctuality of trains, types of services available at the terminals or railway stations and any indicators of safety associated with the maintenance of railway tracks and signals and with overcrowding and congestion.

An important indicator of quality in service industries is the technical quality of service. In the case of industries like telecommunications, electricity, gas and water, the quality of service may be affected by the outage, time taken to undertake repairs, and the promptness and friendliness of the staff in attending to repairs of faulty telephone equipment and lines. Technical quality in the provision of services requires significant outlays on the design and maintenance of equipment and networks used in the service provision.

Measuring quality in health services is a complex problem. The number of complaints from patients, number of hospital days required for a particular health problem or surgical procedure, and cases of negligence could be used as useful indicators of quality of service. In sectors like health and education, quality of service provision may be associated with labour rather than the physical capital. The quality of medical practitioners and nursing staff are crucial to medical services of a hospital. In addition to the provision of modern equipment, staff may need to improve their skills through regular training.

Once there is recognition that considerable variation in the quality of goods and services exists, the question is how we account for quality differences in productivity studies. There are two possible options suggested here.

- The first option is to incorporate the quality differences directly into the output measures, i.e., one could attempt to derive quality-augmented output measures. Such adjustments for quality in output measures are more readily feasible in the case of physical outputs of commodities like TV sets through the use of hedonic regression methods. One commonly used approach is to use nominal value of the total output as the starting point and then adjust the value aggregate using an appropriate price deflator that takes quality differences through an appropriate hedonic regression model. We advise the reader to consult Triplett (2004), *Handbook on Hedonics*, for a comprehensive treatment of the methods used in constructing quality adjusted price index numbers.

- In some cases it may be possible to accord some numerical weights to outputs of different qualities. One should use an objective approach to identify weights. For example, in the case of training university students, students in different disciplines could be weighted according to the relative costs of training such

students. One should refrain from using arbitrary weights that would render efficiency analysis very subjective.

- The third option is to apply a two-stage approach to account for differences in quality of output. In the first stage, unadjusted output measures are used for purposes of productivity analysis. Results from such an analysis are obviously affected by differences in quality of service. For example, a firm may look more efficient because it services more customers per a given level of input but this may be at the cost of low quality of service. So, in making comparisons of performance, measures of efficiency and productivity derived at the first stage may be adjusted for quality in the service provision through an appropriate econometric analysis. We suggest the use of regression approach where efficiency scores or measures are regressed against a variety of quality characteristics and check if quality differences can adequately explain differences in performance.

- Another option is to include quality characteristics directly in the method that is used for computing or estimating technical efficiency scores. It may be possible to use models similar to that used in Battese and Coelli (1995).

In summary, quality variation in outputs is an important issue that deserves careful consideration in productivity studies. It is important that an attempt is made to account for variations in the quality of goods produced or services provided.

5.3 Inputs

A commonly-used classification of inputs involves five categories: capital (K); labour (L); energy (E); material inputs (M); and purchased services (S). The construction and use of data according to these categories in productivity measurement is sometimes referred to as the KLEMS approach. Often, the last three categories of inputs are aggregated to form a single "other input" category. Even though a classification of this type appears fairly intuitive, matters could be more complex in real life. For example, suppose a firm outsources all its information technology (IT) needs such as electronic processing of data, preparation of payrolls and other services for a monthly fee. It is quite obvious that the outlay of the firm on IT services is an important input, but it is not clear to which category it belongs. IT services require capital (in the form of computers), labour in the form of analysts and material inputs, such as paper, ink and other consumables, and energy in terms of power used. Similarly, if, on a long-term basis, a company hires heavy machinery rather than purchasing it, how should this variation in the company practice be accounted for?

Labour

Labour constitutes a major component of the total expenditure on inputs in many enterprises. Labour and capital are the two primary inputs of considerable importance. Despite its importance, usually very little attention is devoted to the measurement of labour. This may, in practice, be due to the apparent ease in measuring this input. However, a lot of important problems in measuring labour input are often ignored or are not explicitly recognised.

The quantity of the labour input is normally measured using a single aggregate variable. The most commonly-used measures of labour input are:

- number of persons employed;
- number of hours of labour input;
- number of full-time equivalent employees; or, in some instances, simply
- the total wages and salaries bill

For analysis at the enterprise level, more accurate measures of the labour input are usually available. Number of employees is a commonly-available measure. Sometimes staff numbers are categorised into full-time and part-time, in such cases, it is necessary to have data on the extent of involvement of part-time employees in constructing a measure of full-time-equivalent employees in a particular enterprise. This is a measure that is very commonly available.

In many cases, estimates of the number of hours worked are also commonly available. Number of hours worked is a more accurate and preferred measure of labour input as it takes into account as to whether a person is employed on a full-time basis or a part-time basis and also the number of hours worked by a full-time employee.

If total wages and salaries bill is used as a measure of labour input then it is necessary to make adjustments for differences in wage and salary levels faced by different enterprises. For example, two different firms located in two different locations or cities may have to pay different wage rates. This is also the case when the firms are located in metropolitan areas or in some rural or remote areas. We note here that wages bill is also affected by the composition and quality of labour, this issue is discussed in detail.

If we have to choose or recommend an appropriate measure of labour input, total number of hours worked is the best indicator of labour input. Where labour is classified into a number of different types, it may become necessary to derive an aggregate measure along the lines discussed below.

Composition and quality of labour

An aggregate or summary measure, such as the number of full-time-equivalent employees, ignores the skill-level composition of labour. A few types of categories commonly distinguished in labour force statistics are listed below.

- Skilled and unskilled labour: Usually skilled and unskilled classification is based on the educational qualifications required to undertake a job. Some jobs require professional qualifications like a degree in engineering or medicine and some other positions like academic positions in universities require a postgraduate tertiary qualification.
- Salaried or non-production or administrative staff and wage or production employees: Employees may be classified as *blue* or *white* collared workers and in such cases it is possible to measure labour input under these two general categories.
- Data on employees classified by age, gender and education levels provides more detailed information that can be used in making appropriate adjustments for quality differences.
- employees classified by their qualifications, e.g., nurses and doctors.

Accounting for the differences in the skills of workers is quite important in measuring the labour input into a production process. For example, Jorgensen *et al.* (1987) have meticulously measured the labour input by taking into account quality in defining characteristics like age, education, class of workers, occupation and gender and constructed these measures by detailed industry categories.

It is much easier to account for different types of labour when data for different firms within an industry are being analysed. For example, if aged-care facilities are being benchmarked then it is quite likely that all the enterprises in the industry employ similar categories of labour, e.g., skilled and unskilled nurses, doctors and other labour. In such cases, all that is needed is to appropriately aggregate employee numbers in different categories using their share in the total wage bill. A way of handling the differences in composition is to make use of an index number approach.

For example, let L_{ik} represent the amount of labour of type k ($k = 1,2,\ldots,K$) used in the i-th enterprise or firm. Let v_{ik} represent the value share of the k-th type of labour in the i-th enterprise. It is possible to construct an index of labour use across different firms using the approaches described in Chapter 4, by first making bilateral comparisons across firms i and j using a Tornqvist index, denoted by L_{ij}^{TT}, as:

$$L_{ij}^{TT} = \prod_{k=1}^{K} \left[\frac{L_{jk}}{L_{ik}} \right]^{\frac{v_{ik}+v_{jk}}{2}} \tag{5.1}$$

where v_{ik} and v_{jk} are the shares of k-th type of labour input in the total labour costs in i-th and j-th enterprises.

We note that the bilateral index in equation (5.1) is not transitive and hence we recommend that a transitive index, using the EKS procedure that is discussed in Chapter 4, be used in arriving at a consistent set of comparisons of labour inputs across firms. This is given by:

$$L_{ij}^{EKS} = \prod_{\ell=1}^{I} \left[L_{i\ell}^{TT} \bullet L_{\ell k}^{TT} \right]^{1/I} \tag{5.2}$$

where I is the total number of enterprises. Selecting one of the enterprises as the base, equation (5.2) provides an index of labour input across all the enterprises. This index is preferred to any physical measures such as the simple numbers of employees in different categories.

For purposes of productivity growth calculations over time, the formula in equation (5.1) can be employed over two time periods, rather than across firms, and the resulting index can be used as a measure of the aggregate labour input. Some comments are in order here:

- One important feature to note is that, in order to construct these indices, only data on the cost share of various labour categories are needed. Usually such data are available from the accounts maintained by the enterprises.
- In some instances, only estimates of the total wage bill are available for each enterprise. In this case, it is advisable that the wage bill is deflated by an index of wage costs that is appropriate for the industry either over time or across regions within a country. If the wage bill is not appropriately adjusted for differential wages, one may have a biased measure of labour input which can in turn lead to biased measures of efficiency and productivity performance.
- If an appropriate index of costs is not available, it is necessary to construct an index using industry-specific wage rates and some estimates of shares of different types of labour. As labour costs usually constitute a major portion of total input costs, it is important that a good measure of labour input is constructed and appropriate wage cost deflators are constructed for purposes of obtaining a reliable indicator of labour input.

Capital input: capital stock, capital service flows and user costs

A proper measurement and treatment of capital in efficiency and productivity studies is very important. This is particularly significant in cases where multi-factor productivity measures are sought and proper account of all the inputs used in the production process is taken. The measurement of quantity and price of capital input is quite difficult and challenging. The main reason for the difficulty in the treatment

of capital is that it is a durable input. Unlike material inputs or labour input, which are consumed or utilised in the production process within an accounting period, capital assets are purchased in one period and used in the production process through the life of the asset or until it is replaced by a new asset. Several questions are important here. How does one measure the total capital stock of a firm? Do we use capital stock or the flow of capital services as a measure of the capital input? What is the price associated with the capital input? In this short exposition, it is difficult to address all these issues, but the reader is referred to the OECD Productivity Manual (OECD, 2001a), OECD (2001b) Manual on Capital Stock Measurement, and a few other references that are devoted to the measurement of capital, including Christensen and Jorgensen (1969), Diewert (1980), Diewert and Lawrence (1999), Griliches (1963), Harper *et al.* (1989), Hulten (1990), Hulten and Wykoff (1996) and Jorgensen (1993).[1] Chapter 5 of the OECD manual also provides worked examples of calculations involved in deriving estimates of capital stock and in the calculation of user costs.

Typically, the capital input used by a given enterprise may be measured by the total service flows from various capital assets of the firm. Assets may refer to buildings, tractors, computers and other equipment that has the potential to provide services over a period of time. In assessing productivity measurement, increases (or changes) in the volume of capital service flows used over time or differences in levels across different firms are considered for each asset type. These are then aggregated to determine the use of capital services in the production process. It is possible to identify a few major steps in the determination of the capital input use.

1. Determination of productive stock of a capital asset

Obtaining an accurate measure of productive capital stock by asset type is a crucial step in determining the flow of capital services used in the production process. However estimates of capital stock are not always readily available. In instances where we are interested in measuring productivity growth over time, it may be possible to use investment series to derive estimates of productive stock. But in the case of productivity comparisons over cross-sectional units, one may need to use a pragmatic approach and make use of data that are readily available. In such instances one may consider measuring quantity of capital using estimates of replacement value or sales values of assets. We discuss both of these options briefly. First we start with the more commonly used *perpetual inventory method* and then discuss more practical ways of measuring capital stock.

The Perpetual Inventory Method (PIM)

The construction of the productive stock by asset type is the first step in determining the use of capital services. Assets may be classified broadly into buildings, small

[1] This list is only a small selection from numerous studies devoted to this subject. For a more complete list the reader is referred to OECD (2001a).

machinery, heavy machinery, transport equipment, computers and other types of assets. The classification used should be relevant to the particular problem under consideration. If firms producing TV sets are under consideration, buildings, small machinery used in the assembly, machinery for final packaging and transport equipment may be relevant. In contrast, if we consider productivity of universities then buildings, library collections, computing and information technology equipment, transport vehicles may be more relevant. Once a list of capital assets is prepared then one can set about the construction of capital stock for each of the asset categories.

Measurement of capital stock is usually based on the *perpetual inventory method* (PIM), which requires the following data for purposes of measuring capital stock by assets.

- A time series of investment expenditure on the particular asset over a long enough period (depends upon the productive life of a given asset). The type of assets included in the class should be narrow enough so that a single lifetime can be representative of all the components.

- The second piece of information needed for the computation of productive stocks is to produce price index numbers of investment goods to deflate the investment expenditure series. The deflated expenditure series provide a series of real investment expenditure. Availability of an appropriate series is quite crucial here. For example, it is quite difficult to find appropriate deflators for items like computers where price and quality changes are rapid. Particular attention should be paid to the choice of deflators for computing equipment and information technology related products.

- The third piece of information needed is the retirement patterns for different assets. Retirement patterns depend upon the service life as well as the assumptions made about the pattern of service flow around this service life or the retirement patterns. Service life of an asset refers to the length of time that assets are retained in the capital register. Estimates of asset lives need to be derived from company records, through surveys or from the information obtained from the manufacturers and, finally, using expert advice. Choosing a longer life length results in an increase in the size of the gross capital stock for a fixed investment series and reduces the consumption of fixed capital. This, in turn, increases the size of the net capital stock.

- The next piece of information required is the retirement pattern of the asset. Some of the commonly used patterns are: linear, delayed linear, bell-shaped, simultaneous exit and Winfrey mortality functions. For an illustration of the shape of these functions, see OECD (2001b).

These four sets of information are useful in constructing measures of the gross capital stock. Following Appendix 4 of the OECD (2001a) *Productivity Manual*, we can provide a formula for the computation of capital stock estimate using the perpetual inventory method.

Suppose K_t^P represent the productive capital stock of a particular asset under consideration. Then it is given by:

$$K_t^P = \sum_{\tau=0}^{T} h_\tau \cdot F_\tau \cdot \frac{IN_{t-\tau}}{P_{t-\tau,0}}$$

where IN_t is nominal (current price) investment expenditure on the particular asset in period t; $P_{t,0}$ is the price index for the asset in period t with period 0 as the base; F_t is a retirement function showing the share of assets aged τ that are still in service (taking values between 1 and 0); h_t is an age-efficiency profile tracing the productive efficiency of the asset (varies between 1 for a new asset to 0 for an asset which has lost its entire productive capacity); and, finally, T represents maximum service life of an asset.

In order to empirically implement the formula, we need data on historical investment series going back at least up to the service life of an asset. So if an asset has a 20 year service life, we need investment series for the past 20 years to be able to have a capital stock estimate in the beginning year. We also need a good price index to deflate the investment series, the price index should take adequate account of the change in quality of the asset over time.

This can be considered as an intermediate step in estimating the productive capital stock that also takes into account the age-efficiency of the asset over its lifetime.

- The final data required is the age-efficiency pattern of the productive asset. These patterns make it possible to account for the loss of productive capacity of the asset as it ages. The age-efficiency pattern reflects the wear and tear of the assets. Some of the commonly-used patterns are:

  (i) one-hoss-shay efficiency profile, which assumes that as long as the assets exist their productive capacity remains fully in tact and drops to zero when the asset is retired;

  (ii) a straight-line depreciation model where efficiency declines by a fixed amount each year;

  (iii) a declining balance or geometric depreciation model where efficiency falls at a constant rate; and

  (iv) a hyperbolic pattern where productive services of a capital good fall slowly in early periods of life and at an increasing rate in later periods.

The following chart shows the age-efficiency profiles under the four different assumptions.

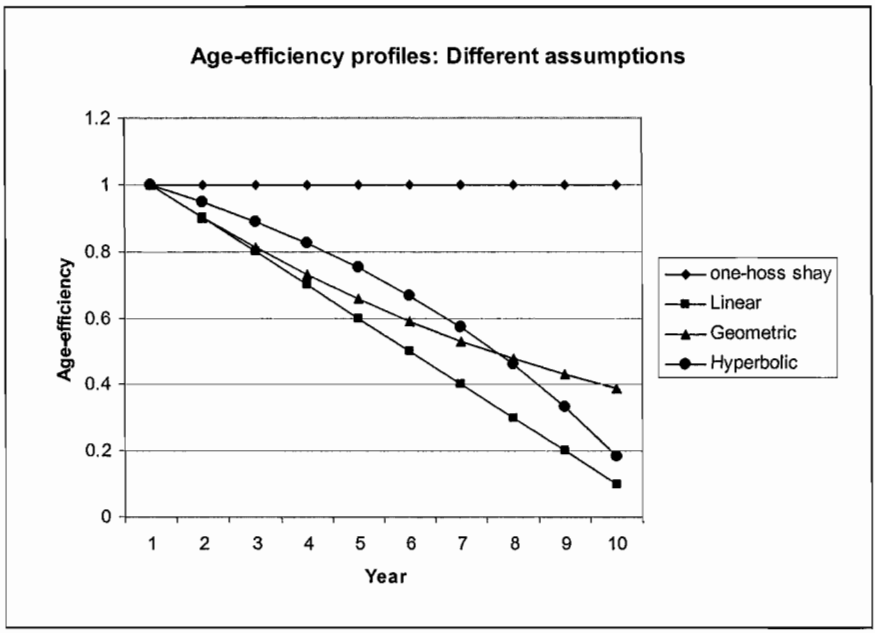

Figure 5.1 Age efficiency profiles under different assumptions

The linear profile assumes that efficiency decreases by a fixed 10 percent of the capacity each year; geometric profile assumes that efficiency loss in any given year is 10 percent of the efficiency at the beginning of the year; and the hyperbolic profile is given by:

$$E_t = E_0 \left(T - (t-1)\right) / \left(T - \beta(t-1)\right)$$

where β is the slope-coefficient which is set at 0.5 in the above Figure. It can be seen that the hyperbolic function portrays a slower decline in efficiency in the beginning years and a more rapid decline in the later parts of the asset life. It is exactly opposite of what happens in the case of geometric pattern of efficiency life of the asset.

Diewert and Lawrence (1999) consider a number of efficiency profiles and obtain estimates of capital stock for Canada using alternative assumptions. OECD (2001a) provides simple worked examples that are useful in understanding the underlying concepts. The Australian Bureau of statistics uses a hyperbolic profile with a value of β equal to 0.5. Careful consideration should be given to the choice of the type of profile and then an appropriate value of the parameter involved needs to be chosen.

Alternative measures of quantity of capital

In the discussion thus far, we have assumed the availability of data required at each stage of the calculations. We need data on investment history as well as appropriate price index numbers for the capital assets involved. In situations where we do not have investment history data, a few alternative measures of the quantity of capital may be considered. This material is drawn from Coelli *et al.* (2003). These are described briefly below.

Replacement Value

An undepreciated replacement value of capital stock held by a firm should, in theory, be equal to the undepreciated value of capital stock in constant prices. Estimating the replacement costs of all items of firms amounts to a large-scale survey of the assets. However, such a survey needs to be undertaken only once and from then on it may be possible to update the capital stock using investment data from the time the survey is conducted.

Sale price

This is market price obtained from the sale of an enterprise. One can argue that the value of the capital stock should be reflected in the sale price. However, it may not always be possible to disentangle the price of the capital stock out of the total sale price. Since sales of a given type of enterprise are not always that common, estimates of capital stock from sale prices could be unreliable.

Physical measures

If capital stock estimates by assets are more difficult to arrive at, it is possible to make use of some physical measures or proxies. Depending upon the particular industry to which firms belong, it may be possible to classify capital into broad categories and identify some simple measures. For example, capital items could be classified into buildings, small machinery, heavy machinery, vehicles, computers and others. Under each category one can find some physical indicators. For example, total floor area could be used as a proxy for buildings; number of desktop computers, total horsepower of machines of a particular type and number of cars can be used indicators for other capital items.

There are always problems associated with the use of physical measures. Variation in quality of the indicators is a major problem. For example, aggregating different types of computers into one category or using number of cars where cars can be of varying size with different features is always a problem.

When some alternative physical proxies are available, it is a good strategy to examine the sensitivity of the productivity measures to the choice of the proxies. We

recommend that some checks for robustness of the results be conducted before a measure is selected and before the results are finalised.

Other measures

Other measures of capital stock include the undepreciated and depreciated nominal capital stock. Such values are routinely reported in annual accounts of the enterprises. An alternative measure that is based on the depreciated replacement value could provide a better measure of capital stock. See Coelli *et al.* (2003) for a worked example of a variety of these alternatives.

It should be emphasised that these alternative measure should be used only when data are limited and one is forced to use accounting records to get estimates of capital stock for purposes of productivity measurement.

2. Measurement of capital services

Measuring the actual flow of capital services is a more difficult task. The productive service of an asset is typically assumed to be a proportion of the productive capital stock of the asset. Hence, the productive capital stock reflects the quantity of capital services used in production. This assumption forms the basis for the use of measures of capital stock as an indicator of the use of capital services. If the factor of proportionality does not change over time (or across firms) then the growth rate of capital services (differences in levels across different firms) is identical to the rate of growth of the capital stock. This is clearly an unrealistic assumption as the level of utilisation of the capital stock may vary over time. Here, the rates of capacity utilisation of an asset over time become important. In cases of business cycles of economic activity, the proportionality assumption produces cyclical swings in the measure of total factor productivity that includes capital services as an input.

Obviously, a satisfactory treatment of the varying degrees of the proportionality between capital stock and the service flows is crucial. There have been several attempts to measure rates of capacity utilisation and a few are discussed in OECD (2001). In the absence of reliable measures of capacity utilisation, it is more sensible to compare productivity performance at similar phases of the cycle. However, such an approach is of limited use in the context of cross-sectional productivity comparisons.

3. Determination of user costs of capital

Capital services are the flow of services from capital goods into the production process, and so their value can be considered as an appropriate measure of quantity of capital used. Then what is the price attached to this quantity? The price component is measured using the *user cost* or *rental price* associated with the use of a particular asset. Following OECD (2001), the user cost of an asset whose market price of the new asset is q_t is given by:

$$\mu_t = q_t \cdot (r_t + d_t) - (q_t - q_{t-1}) \tag{5.3}$$

where μ_t is the per-period cost of using the services of the asset; q_t is the market price of a new asset; d_t is the rate of depreciation; and r_t is a measure of the cost of financial capital such as the market rate of interest.

The first term in the expression above shows the cost of financing the asset and the second component measures the capital gains or losses or the revaluation of an asset. This represents the rise or fall in the value of the asset independent of the effects of ageing. It is quite important to measure the rates of depreciation and the returns on financial capital as they affect the estimate of the user cost of capital.

4. Aggregation across assets

All the measurement discussion so far focused only on a single asset. Typically, the capital input into a production process involves a number of productive capital assets. Therefore, there is a need to aggregate the capital service flows across assets. Usually aggregation is undertaken in the form of an index. Thus, an index of capital service flows over time or across firms or enterprises is constructed. The following formula provides an aggregation scheme under the assumption that capital service flows are proportional to the capital stock.

Let $K_{i,t}$ and $\mu_{i,t}$ respectively represent the productive capital stock and the associated user cost of the i-th asset in period t. Then, it is possible to use a Fisher or Tornqvist index in deriving the quantity index of aggregate capital services. The following formula is based on the Tornqvist index:

$$T_{t-1,t}{}^{K} = \prod_{i=1}^{N} \left[\frac{K_{i,t}}{K_{i,t-1}} \right]^{\frac{v_{i,t} + v_{i,t-1}}{2}} \tag{5.4}$$

where $v_{i,t} = \dfrac{\mu_{i,t} \cdot K_{i,t}}{\sum_{i=1}^{N} \mu_{i,t} \cdot K_{i,t}}$ where N represents the number of capital assets used in the aggregation.

In the case of comparisons across firms, the Tornqvist index defined in equation (5.4) needs to be modified and an EKS approach be used for the construction of transitive quantity indices.

Once all these steps are completed, we have an adequate measure of the flow of capital services into the production process. This is essential if we aim to arrive at a proper measure of total factor productivity trends and levels.

Energy, materials and purchased services

This is a very important category of inputs. Energy and material inputs account a significant share in input costs in particular sectors like agriculture and iron and steel plants. In practice, these three types of inputs are aggregated into one category consisting of "other" inputs. However, it is usually very easy to obtain both quantity and price data for the energy inputs. Material inputs usually consist of all intermediate inputs drawn from other sectors of the economy. Usually, outlays on material input information may be available in considerable detail from the accounts of the enterprises. However, it is not always possible to use information in such detail. Hence, it becomes necessary to aggregate them into one or two consolidated categories.

Expenditure on material inputs is available in nominal form. We need to make use of an appropriate deflator to obtain real expenditure that can be used as a measure of "quantity" of material inputs. In the case of cross-sectional analysis where efficiency and performance of firms at a single point of time are being analysed, it is necessary to make an adjustment for spatial differences in prices. However, in practice such deflators are not commonly available. In this case the performance measures derived using nominal values need to be interpreted with caution.

Purchased services and outsourcing

Expenditure on purchased services is often considered as an intermediate input and data on this item are usually aggregated with other material inputs in the accounts of the firm. In the past, when purchased services had a negligible cost share, treatment of purchased services may not have been a major problem. But, in recent years, there has been an increasing tendency to outsource a number of services such as cleaning, security, and computing and IT-related services. Outsourcing offers flexibility to the managers of enterprises and allows them to respond to market forces more efficiently. There are instances where several firms may share a common administrative unit. For example, a religious organisation which runs schools or nursing homes in different locations could share administrative services from a centralised location.

It is quite possible to treat this as a separate input item and make efforts to come up with an appropriate quantity and price measure that can be used in productivity measurement. However, there are a few conceptual problems. For example, outsourcing a fairly labour-intensive activity would mean that the enterprise could cut its direct labour use and increase its "other input" expenditure. Such a shift in input expenditure has two immediate effects. First, it has the tendency to show marked gains in labour productivity even though the overall use of labour may not have decreased, but has just shifted into another item of input expenditure. Second, inclusion of purchased services in the "other input" category increases the size of the "other input" relative to those firms that do not outsource some of their

requirements. In such cases, outsourcing firms may appear to be more productive than the others.

Therefore, it is necessary to treat outsourcing expenditures carefully. Where possible, one should reallocate components of such outlays into more traditional input categories such as labour, capital services and material inputs. If such a reallocation is not possible, then one should be cautious in the treatment of those enterprises in any benchmarking exercise and measures of performance should be appropriately interpreted.

5.4 Prices

From the last two sections, it is abundantly clear that price data play a major role in productivity measurement, either directly or indirectly. Price data are used for purposes of aggregation, leading to appropriate measures of real output and real expenditure on various inputs. Price indices of major output groups and input categories could be used just like standard price data in the econometric estimation of cost and profit functions and also in assessing allocative efficiency of firms under consideration.

It is standard practice that prices, for outputs as well as inputs, represent producer prices. That is, output prices are prices received at the farm or factory gate and should exclude transport costs and marketing margins. It should include any subsidies received by the producer directly for each unit of output produced. In contrast, input prices should include prices paid for inputs that include all the taxes and marketing margins incurred in the purchases of the inputs.

A major problem concerning price data is that firms do not often record prices paid for the inputs purchased. Often total quantities of inputs purchased for the year may or may not be recorded but almost always the total expenditure on each input category is carefully recorded in the accounts kept by the firm. Similarly, total revenue and total quantities sold of each commodity produced by the enterprise may also be recorded. In both of these cases, it is possible to derive "unit values". These unit values are essentially average prices of a broadly-defined group of commodities that may exhibit some quality variation across commodities. For example, unit value for rice would be an average price for a range of rice qualities that were produced and sold by a farmer. Unit values may also be derived for inputs such as unskilled and skilled labour. These are obtained by simply dividing the total wage bill for unskilled and skilled labour, respectively, by the full-time-equivalent number of staff under these two categories. These average wage rates are an amalgam of differential prices, as well as differential composition of labour, in terms of skills possessed. The use of actual wage rates associated with different types of labour is ideal if data exists for detailed categories.

Another issue concerns the nature of the price data. Are the prices essentially market prices that directly influence the behaviour of the producers and consumers? Or are the prices essentially administered prices, in which cases prices could be distorted and may not have the same economic meaning with respect to producer or consumer behaviour

In the absence of direct price data, one has to rely on the choice of an appropriate price deflator. Too often, when conducting empirical studies using sophisticated econometric techniques, researchers are quite happy to use whatever deflator they may be able to find. From the index number theory and practice described in Chapter 4, it is clear that the deflator must be as close as possible to the prices of the group of commodities. This is essential if we wish to interpret the deflated value as a real aggregate or quantity. Usually, national statistical offices are an excellent source of data for such deflators – they produce a range of index numbers including the consumer price index, producer price index, GDP deflator, import and export price index numbers, wage cost index, etc.. These publications, not only provide a measure of the headline index, but also publish indices for subcategories. Researchers should devote adequate time to find the right deflator for the aggregate under consideration.

We draw the attention to an important data problem with respect to appropriate deflators for purposes of adjusting nominal aggregates for firms, located in different towns and regions of a country, at a given point of time. Most deflators available refer to temporal movements in prices. At a given point of time, prices may indeed exhibit more variation that is shown in temporal movements of prices. Reasons could relate to transport costs, market rigidities and other factors that influence prices especially wages. If no adjustment is made of these differences in levels of prices, and if nominal aggregates are taken as a measure of the underlying quantities, then significant errors could creep into productivity comparisons. At this time, only deflators for cross-country comparisons are available, not for regions or cities within a country. One must be aware of this problem and use any information that could improve the underlying quantity measures.

5.5 Comparisons over Time

Our discussion thus far has mainly focused on data issues concerning for comparisons across firms and enterprises at a given point of time. Now we briefly focus on comparisons over time. All the measurement issues discussed in Sections 5.2 to 5.4 are also relevant for temporal comparisons, but a few additional issues arise. There are two issues of particular concern. First, there are observed movements in prices of all output and input commodities and services. Therefore, it becomes necessary to make appropriate adjustments for price changes before embarking on actual productivity measurement. Second issue concerns the issues of comparability that arise due to changes in the quality of products and due to new and disappearing goods. It is important that these issues are adequately addressed.

Adjusting for price changes over time

It is possible to make adjustment for price changes after constructing appropriate price index numbers. We draw particular attention to the following issues.

- *Choice of formula:* From the exposition on index number theory and practice, use of Fisher or Tornqvist index number formula to measure price changes is the most appropriate. Both of them have several attractive economic theoretic and test properties. You need to recognise that if you are using published price indexes such as the consumer price index (CPI) then it is more than likely that such an index is derived using the Laspeyres index.

- *Fixed base versus chain base indexes:* If you are working with a long time-series data, it is appropriate that you make use of a chain index. In any case it is important not to use an index that makes use of base period which is too far from the current period. For example, if you are working with data over the period 1970 to 2000, use of fixed-base price indexes that use 1970 as the base period should be avoided. It is important that chained price indices at least at 5 or 10 year intervals are used.

- *Which deflator to choose?:* In practical situations, you may not have detailed price and quantity or value share data to construct your own price index numbers. In such instances you need to make use of published index numbers. The problem here is that you may not find a price index that matches your exercise. In this case you need to select a price index number that suits your problem the best. In too many productivity studies, researchers are happy to make use of whatever price index they can get access to. It is quite common to see that researchers just make use of the consumer price index (CPI) or an implicit gross domestic product (GDP) deflator drawn from national accounts data. From the discussion in Chapter 4, it is quite clear that *the deflator selected must relate to the commodities that constitute the aggregate as closely as possible.* It is important to remember that, to the extent possible, the domain of the price index should be the same as the domain of the value aggregate. It is only then the deflated value aggregate, or value aggregate at constant prices, can be considered as a volume or quantity measure. If you use a wrong deflator, then you will have inappropriate quantity measures which will affect your analysis. So using CPI may be quite inappropriate in some cases, such as in the case of deflating the nominal output of firms in the textile and clothing industry. The researcher must sift through all the data sources and select a deflator that is best suited for the purpose.

- *Adjustments for quality change:* When we consider the case of multiple products, a few other data problems may arise. When analysing productivity performance over time, the problem of new and disappearing goods is usually encountered. A related problem concerns rapid quality changes over time, as is the case for firms

producing computing equipment. Usually hedonic regression methods are used in constructing the price index numbers used for deflating value aggregates. OECD is currently preparing a manual, (Triplett, 2004), outlining procedures for the construction of quality adjusted price index numbers. ILO (2004) is also a useful manual on the construction of the consumer price index. It has been found that the use of price index numbers for computing equipment which does not adequately account for quality improvements can seriously understate the computing or information technology input into the production process thereby resulting in exaggerated estimates of productivity change. If you have IT related equipment expenditure then you need to make sure you make use of a quality-adjusted price index.

5.6 Output aggregates for Sectoral and Economy-wide Comparisons

Productivity measurement and comparisons across regions within a country and over time are commonly undertaken at an aggregate level. In most growth and development studies, researchers are interested in the productivity growth performance of a sector, like agriculture or transport, or on the whole economy. Sectoral analysis of productivity is usually based on value aggregates and the usual approach of identifying capital, labour, energy, materials and service inputs does not work very well due to the sector nature of the entities involved. A few interesting problems and issues arise in this context that a researcher should resolve satisfactorily before embarking on any serious analysis of the data.

- We make a general comment here on the form in which data are available for sectoral comparisons. Data are usually in a time-series format. For example, agricultural sector output and input data are available either on an annual basis or on a quarterly basis. Thus productivity growth estimates are based on just time-series data. In this case we recommend the use of the index number approach. For this approach, we need to measure quantities (of output and input goods and services) and their prices.

- The first and foremost problem is to choose an appropriate measure of output. For example, what is the output of the manufacturing sector? Two concepts are often used: (i) Gross output, which is measured as the sum of the value of outputs of all the firms belonging to the sector. and (ii) Gross value added, which is a measure of the total value of output net of all the sectoral output that is used as an intermediate input (non-labour and non-capital inputs) into the sector itself. Thus, gross value added may be considered as the contribution of a given sector to the economy. Use of either of these aggregates leads to a different measure of output and inputs used in the production process of the sector – care must be taken to adjust inputs to be consistent with the output measure used. In theory, both of these approaches can be used. Balk (2004)

discusses the analytical link between measures of productivity based on gross output and gross value added.

- Gross value added data are more commonly used when sector analyses are undertaken. National accounts publications are the main source of such data.

- Appropriate deflators should be used in constructing real aggregates for use with productivity measurement techniques. Again, national accounts publications are a good source. Most national accounts publications provides estimates of GDP with a sectoral breakdown, viz., agriculture, mining, manufacturing and the services sector, at current and constant prices. These can be used in deriving appropriate price deflators. In selecting a deflator, make sure it corresponds as closely as possible to the aggregate you are working with.

5.7 Cross-country Comparisons of Productivity

In principle cross-country comparisons can be considered like cross-sectional units like firms or enterprises except that we have data for each country as a whole. Studies of catch-up and convergence and productivity performance of countries are quite common. So we make a few comments on data related issues when international comparisons of productivity are undertaken. The following issues need to be considered.

- First, we reiterate the point that cross-country comparisons are usually undertaken at an aggregate level. So researchers should make use of appropriate value aggregates following the comments made in the previous section. Comparisons of GDP or sectoral value added are common practice in this connection.
- An additional problem is encountered when value added or GDP data are used for purposes of cross-country productivity comparisons (see, e.g., Färe *et al.,* 1994; Rao and Coelli, 2002; and Coelli and Rao, 2004). In these cases, published data are usually expressed in national currency units. For example, GDP of India and Australia are in rupees and Australian dollars, respectively. So it becomes necessary to convert these into a common currency unit. A number of alternative currency converters are available for this purpose:

    - Exchange rates (source: *The International Monetary Fund)*
    - Purchasing power parities (PPPs) from the International Comparison Program (ICP) (source: *The World Bank*)
    - PPPs for different sectors like agriculture, manufacturing and services (source: *Groningen Growth and Development Centre; FAO)*

- Out of various conversion factors, it is recommended that PPPs are used but care must be taken to ensure that the PPPs refer to the value aggregate used in the comparisons. Though exchange rates were used for international comparisons of

economic aggregates their use in currency conversion can seriously bias the real aggregates derived. It is now widely accepted that exchange rates do not adequately represent the purchasing power of currencies, rather they reflect the demand and supply for a given currency.

- For overall economy comparisons involving comparisons of gross domestic product, it is quite appropriate to use PPPs from the ICP. The ICP provides PPPs for only selected benchmark years and for only those countries that participate in the international comparisons programme. But if we wish to undertake productivity analyses covering a large number of countries and over time, the Penn World Tables is a very valuable source of PPPs and the latest version PWT 6.1) provides a panel data on GDP, consumption and investment covering a large number of countries spanning a 50-year period. PPPs are just like deflators – they take into account both prices and currency conversions at the same time – so it is important that the correct PPPs are employed for the problem at hand.

- Once the output side of international comparisons is taken care of, it is necessary to focus on the input side of international productivity comparisons. The two inputs normally considered here are labour and capital. In some agricultural sector comparisons the land input is also considered (see Rao and Coelli, 2004; Coelli and Rao, 2005).

- Estimates of labour force are usually based on total number of persons employed or the total number of hours worked. Such data are available for most of the developed countries. However, in developing countries such detailed labour force data are not available. In the absence of estimates of labour input, normally total workforce, population in the age 15 to 65 of a country, is considered to be the only choice. For purposes of sectoral analysis, population economically active in a particular sector, like agriculture, is taken to be an adequate measure of the labour input. Choice of a particular measure of labour input depends upon the particular empirical application. In many cross-country studies, it is quite common to simply use the total labour force as a measure of labour input without making an appropriate adjustment for unemployment rates or for the presence of disguised unemployment. The main question here is one of full or partial utilisation of the labour force that is available. In the case of studies involving specific sectors like agriculture, it is more difficult to know the unemployment rates, even if one were interested in making an adjustment and deriving persons employed.

- Estimates of capital stock are usually more difficult to derive. Capital stock estimates at the country level are available for a number of countries – these are available on the website of the Groningen Growth and Development Centre (http://www.ggdc.net/dseries/60-industry.html). The Penn World Tables contain an investment series for each country, adjusted for price movements across countries and over time. The perpetual inventory method can be used in

conjunction with these series in deriving internationally comparable capital stock series.

5.8 Data Editing and Errors

Data collection and editing is an important step in efficiency and productivity measurement. Productivity studies often involve large data sets and it may not be possible to visually check for the presence of errors and outliers that could seriously affect the final results. The three main reasons for the presence of outliers are: (i) typographical errors; (ii) invalid observations; and (iii) unusual observations that are real outliers. The researcher should correct the typographical errors and drop the invalid observations or modify them if the reasons are clear and monitor the unusual observations.

Data editing and checking for outliers is an important task that could have serious influences on final results. The standard procedures used in identifying outliers are essentially statistical. If the observations are expected to follow a normal or a mound-shaped distribution, it is possible to identify outliers using z-scores attached to the observations. A few simple and standard procedures are listed below:

- Check for the presence of outliers using sample means, standard deviations, maximum and minimum values and plots of all the variables. Investigate any suspect observations in more detail.
- Look for zeroes in the data and then see if such values are meaningful. For example, having zero for labour or another important input may indicate some problem that deserves further investigation.
- Compare and check suspect data values with alternative sources, if possible. If some series of price indices show some abnormal increases, examine indices for related series which are expected to show a similar trend.
- Check for internal consistency of data. If accounting data are used, it is possible some items are double counted, so make sure all the entries add up to the total. Similarly, if one has the value aggregates, such as the total wage bill, as well as some quantity data, such as the number of full-time-equivalent employees, it should be possible to cross-check any suspect entries by computing unit values and comparing them with salary or wage rate data that may be available for the sector or industry under consideration
- Check some basic ratios, such as output per unit of labour or capital per unit of labour and plot these ratios for all the firms in the data set. A visual check usually reveals some outliers and further examination may be necessary to establish the real reasons for the presence of abnormal figures.
- Run simple regressions to estimate very basic production functions or distance functions and examine the residuals to check for outliers and for observations that exert a lot of influence on the regression equation.

The results obtained from the application of sophisticated techniques, some are discussed in Chapters 6 to 10, can be influenced by the presence of observations

with measurement error and the presence of outliers. Non-stochastic non-parametric approaches like data envelopment analysis are very sensitive to the quality of data used. If and when some data errors go unnoticed, they may result in some abnormal or counter-intuitive results. Before embarking on a lengthy explanation for such results, it may be useful to go back and check data for those firms showing some abnormal results.

5.9 Conclusions

The main purpose of this chapter is to alert researchers to a number of data-related issues that could have an important bearing on the outcome of the empirical examination of productivity and efficiency. Careful examination of the output and input variables, and establishing an appropriate level of aggregation for the study, are important first steps. Choosing between various output indicators, in the case of service industries, could be particularly challenging. A good understanding of the industry under consideration is a first and foremost requirement. A lack of understanding of the industry under study could lead to serious problems in the analysis. In such cases, it may be necessary to recruit an industry expert to the project.

In most practical situations, data compilation is indeed a complex issue and can often be very frustrating. Lack of suitable data is usually a major problem. In dealing with such situations, it is important to first determine the "target" measure that is being sought. The target measure is the theoretically-ideal measure that is necessary for the analysis. For example, we may want a price deflator for a particular class of inputs. But data on such a target measure may not be available. Then, the strategy is to look for, and strive towards attaining, data that are consistent and come closest to the ideal measure. Where second-best options are used, it is important to discuss the reasons for the choice of such measures and discuss the possible effects of such a choice on the final results.

6. DATA ENVELOPMENT ANALYSIS

6.1 Introduction

This chapter is a pivotal chapter in this book because we begin to describe techniques that can be used to measure firm-level inefficiency. In earlier chapters we have discussed index number methods, which implicitly assume that all firms are fully efficient. In the remaining chapters, we relax this assumption and describe methods that may be used to estimate frontier functions and measure the efficiencies of firms relative to these estimated frontiers.

Frontiers have been estimated using many different methods over the past 40 years. Lovell (1993) provides an excellent introduction to this literature. The two principal methods that have been used are data envelopment analysis (DEA) and stochastic frontier analysis, which involve mathematical programming and econometric methods, respectively. This chapter and the next are concerned with the DEA method, while Chapters 9 and 10 discuss stochastic frontiers.

This chapter is divided into several sections. In section 6.2 we introduce a basic DEA model, in which a constant returns to scale (CRS) technology is assumed, while in Section 6.3, we describe a (more general) variable returns to scale (VRS) DEA model. The DEA models discussed in Sections 6.2 and 6.3 are input-orientated models. In Section 6.4, we describe output-orientated versions of these models. The final section contains some concluding comments, which point us towards Chapter 7 in which a number of additional DEA models are discussed.

6.2 The Constant Returns to Scale DEA Model

DEA involves the use of linear programming methods to construct a non-parametric piece-wise surface (or frontier) over the data. Efficiency measures are then calculated relative to this surface. Comprehensive treatments of the methodology are available in Färe, Grosskopf and Lovell (1985, 1994), Seiford and Thrall (1990), Lovell (1993), Ali and Seiford (1993), Lovell (1994), Charnes et al (1995), Seiford (1996), Cooper, Seiford and Tone (2000) and Thanassoulis (2001).

The piece-wise-linear convex hull approach to frontier estimation, proposed by Farrell (1957), was considered by only a few authors in the two decades following Farrell's paper. Boles (1966), Shephard (1970) and Afriat (1972) suggested mathematical programming methods that could achieve the task, but the method did not receive wide attention until the paper by Charnes, Cooper and Rhodes (1978), in which the term *data envelopment analysis* (DEA) was first used. Since then a large number of papers have appeared, which have extended and applied the DEA methodology.

Charnes, Cooper and Rhodes (1978) proposed a model that had an input orientation and assumed constant returns to scale (CRS). Subsequent papers have considered alternative sets of assumptions, such as Färe, Grosskopf and Logan (1983) and Banker, Charnes and Cooper (1984), in which variable returns to scale (VRS) models are proposed. Our discussion of DEA begins with a description of the input-orientated CRS model because this model was the first to be widely applied.

We first define some notation. Assume there are data on N inputs and M outputs for each of I firms. For the i-th firm these are represented by the column vectors $\mathbf{x}_i$ and $\mathbf{q}_i$, respectively. The $N\times I$ input matrix, $\mathbf{X}$, and the $M\times I$ output matrix, $\mathbf{Q}$, represent the data for all I firms.

An intuitive way to introduce DEA is via the *ratio* form. For each firm, we would like to obtain a measure of the ratio of all outputs over all inputs, such as $\mathbf{u}'\mathbf{q}_i/\mathbf{v}'\mathbf{x}_i$, where $\mathbf{u}$ is an $M\times 1$ vector of output weights and $\mathbf{v}$ is a $N\times 1$ vector of input weights. The optimal weights are obtained by solving the mathematical programming problem:

$$\begin{aligned} &\max_{\mathbf{u},\mathbf{v}} (\mathbf{u}'\mathbf{q}_i/\mathbf{v}'\mathbf{x}_i), \\ &\text{st}^1 \quad \mathbf{u}'\mathbf{q}_j/\mathbf{v}'\mathbf{x}_j \leq 1, \quad j=1,2,\ldots,I, \\ &\qquad \mathbf{u}, \mathbf{v} \geq \mathbf{0}. \end{aligned} \tag{6.1}$$

This involves finding values for $\mathbf{u}$ and $\mathbf{v}$, such that the efficiency measure for the i-th firm is maximised, subject to the constraints that all efficiency measures must be

[1] The notation "st" stands for "subject to".

less than or equal to one.[2] One problem with this particular ratio formulation is that it has an infinite number of solutions.[3] To avoid this, one can impose the constraint $\mathbf{v}'\mathbf{x}_i = 1$, which provides:

$$\begin{aligned} &\max_{\mu,\nu} (\mu'\mathbf{q}_i), \\ &\text{st} \quad \nu'\mathbf{x}_i = 1, \\ &\qquad \mu'\mathbf{q}_j - \nu'\mathbf{x}_j \leq 0, \qquad j=1,2,\ldots,I, \\ &\qquad \mu, \nu \geq \mathbf{0}, \end{aligned} \tag{6.2}$$

where the change of notation from $\mathbf{u}$ and $\mathbf{v}$ to μ and ν is used to stress that this is a different linear programming problem. The form of the DEA model in linear programming (LP) problem 6.2 is known as the *multiplier* form.

Using the duality in linear programming, one can derive an equivalent *envelopment* form of this problem:

$$\begin{aligned} &\min_{\theta,\lambda} \theta, \\ &\text{st} \quad -\mathbf{q}_i + \mathbf{Q}\lambda \geq \mathbf{0}, \\ &\qquad \theta\mathbf{x}_i - \mathbf{X}\lambda \geq \mathbf{0}, \\ &\qquad \lambda \geq \mathbf{0}, \end{aligned} \tag{6.3}$$

where θ is a scalar and λ is a $I\times 1$ vector of constants. This envelopment form involves fewer constraints than the multiplier form ($N+M < I+I$), and hence is generally the preferred form to solve.[4] The value of θ obtained is the efficiency score for the i-th firm. It satisfies: $\theta \leq 1$, with a value of 1 indicating a point on the frontier and hence a technically efficient firm, according to the Farrell (1957) definition. Note that the linear programming problem must be solved I times, once for each firm in the sample. A value of θ is then obtained for each firm.

The DEA problem in LP 6.3 has a nice intuitive interpretation. Essentially, the problem takes the i-th firm and then seeks to radially contract the input vector, $\mathbf{x}_i$, as much as possible, while still remaining within the feasible input set. The inner-boundary of this set is a piece-wise linear isoquant (refer to Figure 6.1), determined by the observed data points (i.e., all the firms in the sample). The radial contraction of the input vector, $\mathbf{x}_i$, produces a projected point, $(\mathbf{X}\lambda, \mathbf{Q}\lambda)$, on the surface of this technology. This projected point is a linear combination of these observed data points. The constraints in LP 6.3 ensure that this projected point cannot lie outside the feasible set.

---

[2] It should be stressed that this LP is solved for each of the I firms in the sample and, hence, each firm is assigned a set of weights that are most favourable to them.

[3] That is, if $(\mathbf{u}^*,\mathbf{v}^*)$ is a solution, then $(\alpha\mathbf{u}^*,\alpha\mathbf{v}^*)$ is another solution, etc.

[4] The forms defined by equations 6.1 and 6.2 are introduced here for expository purposes. They are not used again in the remainder of this chapter. The multiplier form has, however, been utilised in a number of studies. The μ and ν weights provide extra information in that they can be interpreted as normalised shadow prices. We discuss the use of shadow price information in more detail in the next Chapter.

As described in Färe *et al.* (1994), the production technology associated with LP 6.3 can be defined as $T = \{(\mathbf{x},\mathbf{q}) : \mathbf{q} \leq \mathbf{Q}\boldsymbol{\lambda}, \mathbf{x} \geq \mathbf{X}\boldsymbol{\lambda}\}$. Färe *et al.* (1994) show that this technology defines a production set that is closed and convex, and exhibits constant returns to scale and strong disposability. In later sections we consider alternative DEA models that correspond to production technologies that have less restrictive properties, such as variable returns to scale and weak disposability.

A Digression on Slacks

The piece-wise linear form of the non-parametric frontier in DEA can cause a few difficulties in efficiency measurement. The problem arises because of the sections of the piece-wise linear frontier which run parallel to the axes (refer to the Figure 6.1) which do not occur in most parametric functions.[5] To illustrate the problem, refer to Figure 6.1 where the firms using input combinations C and D are the two efficient firms that define the frontier, and firms A and B are inefficient firms. The Farrell (1957) measure of technical efficiency gives the efficiency of firms A and B as $0A'/0A$ and $0B'/0B$, respectively. However, it is questionable as to whether the point A' is an efficient point since one could reduce the amount of input x_2 used (by the amount CA') and still produce the same output. This is known as *input slack* in the literature.[6] Once one considers a case involving more inputs and/or multiple outputs, the diagrams are no longer as simple, and the possibility of the related concept of *output slack* also occurs.[7] Some authors argue that both the Farrell measure of technical efficiency (θ) and any non-zero input or output slacks should be reported to provide an accurate indication of technical efficiency of a firm in a DEA analysis.[8]

Now it can be stated that, for the i-th firm, the (measured) output slacks are equal to zero if $\mathbf{Q}\boldsymbol{\lambda}-\mathbf{q}_i=\mathbf{0}$ and the (measured) input slacks are equal to zero if $\theta\mathbf{x}_i-\mathbf{X}\boldsymbol{\lambda}=\mathbf{0}$ (for the given optimal values of θ and $\boldsymbol{\lambda}$). However, we should note that the *measured* slacks reported by linear program 6.3 need not identify all "true" slacks, in the Koopmans (1951) sense. This can occur whenever there are two or more optimal $\boldsymbol{\lambda}$-vectors for a particular firm. Hence, if one wishes to be sure to identify all efficiency slacks one must solve additional linear programs. For more on this issue, see the discussion of slacks in Chapter 7. For the remainder of this chapter, however, we avoid the issue by only considering simple empirical examples in which all efficiency slacks are identified by the LP in equation 6.3. However, as we

[5]From Figure 6.1 we can see that the technology defined by equation 6.3 has the property of strong disposability. This ensures that the isoquant does not "bend back" and display input congestion. The DEA model can be modified to allow for weak disposability. This is discussed in Chapter 7.

[6] Some authors use the term *input excess*.

[7] Output slack is illustrated later (see Figure 6.5).

[8] Farrell (1957) defined technical inefficiency in terms of the radial reduction in inputs that is possible. Koopmans (1951) provides a more strict definition of technical efficiency which is equivalent to stating that a firm is only technically efficient if it operates on the frontier and furthermore that all associated slacks are zero.

argue in the next chapter, the importance of slacks can be overstated for a variety of reasons.

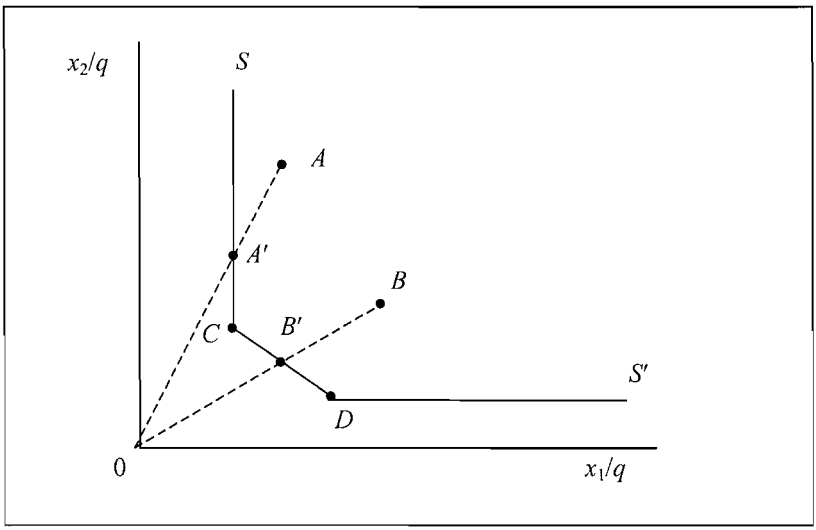

Figure 6.1 Efficiency Measurement and Input Slacks

A Simple Numerical Example

We illustrate CRS input-orientated DEA using a simple example involving observations on five firms that use two inputs to produce a single output. The data are as follows:

Table 6.1 Example Data for CRS DEA Example

firm	q	x_1	x_2	x_1/q	x_2/q
1	1	2	5	2	5
2	2	2	4	1	2
3	3	6	6	2	2
4	1	3	2	3	2
5	2	6	2	3	1

The input/output ratios for this example are plotted in Figure 6.2, along with the DEA frontier corresponding to the solution of the DEA model defined in equation

6.3.[9] It should be kept in mind, however, that this DEA frontier is the result of running five linear programming problems - one for each of the five firms. For example, for firm 3, we could rewrite equation 6.3 as

$$\begin{aligned} &\min_{\theta,\lambda} \theta, \\ &\text{st} \quad -q_3 + (q_1\lambda_1 + q_2\lambda_2 + q_3\lambda_3 + q_4\lambda_4 + q_5\lambda_5) \geq 0, \\ &\qquad \theta x_{13} - (x_{11}\lambda_1 + x_{12}\lambda_2 + x_{13}\lambda_3 + x_{14}\lambda_4 + x_{15}\lambda_5) \geq 0, \\ &\qquad \theta x_{23} - (x_{21}\lambda_1 + x_{22}\lambda_2 + x_{23}\lambda_3 + x_{24}\lambda_4 + x_{25}\lambda_5) \geq 0, \\ &\qquad \boldsymbol{\lambda} \geq \mathbf{0}, \end{aligned} \tag{6.4}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)'$.

The values of θ and $\boldsymbol{\lambda}$ which provide a minimum value for θ are listed in row 3 of Table 6.2. We note that the TE of firm 3 is 0.833. That is, firm 3 could possibly reduce the consumption of all inputs by 16.7% without reducing output. This implies production at the point denoted 3′ in Figure 6.2. This projected point, 3′, lies on a line joining points 2 and 5. Firms 2 and 5 are therefore usually referred to as the *peers* of firm 3. They define where the relevant part of the frontier is (i.e., relevant to firm 3) and hence define efficient production for firm 3. Point 3′ is a linear combination of points 2 and 5, where the weights in this linear combination are the λs in row 3 of Table 6.2.

Many DEA studies also talk about *targets* as well as peers. The targets of firm 3 are the coordinates of the efficient projection point 3′. These are equal to $0.833\times(2,2) = (1.666,1.666)$. Thus firm 3 should aim to produce its 3 units of output with $3\times(1.666,1.666) = (5,5)$ units of the two inputs.

One could go through a similar discussion of the other two inefficient firms. Firm 4 has TE = 0.714 and has the same peers as firm 3. Firm 1 has TE = 0.5 and has firm 2 as its peer. You will also note that the projected point for firm 1 (i.e., the point 1′) lies upon part of the frontier that is parallel to the x_2 axis. Thus it does not represent an efficient point (according to Koopmans' definition) because we could decrease the use of the input x_2 by 0.5 units (thus producing at the point 2) and still produce the same output. Thus firm 1 is said to be radially inefficient in input usage by a factor of 50% plus it has (non-radial) input slack of 0.5 units of x_2 (per unit of q). The targets of firm 1 would therefore be to reduce usage of both inputs by 50% and also to reduce the use of x_2 by a further 0.5 units. This would result in targets of (x_1/q=1, x_2/q=2), which are the coordinates of point 2.

[9] The piece-wise linear frontier isoquant depicted in Figure 6.2 is an example of a unit isoquant. That is, it depicts the various (technically efficient) input combinations that can be used to produce a unit of output.

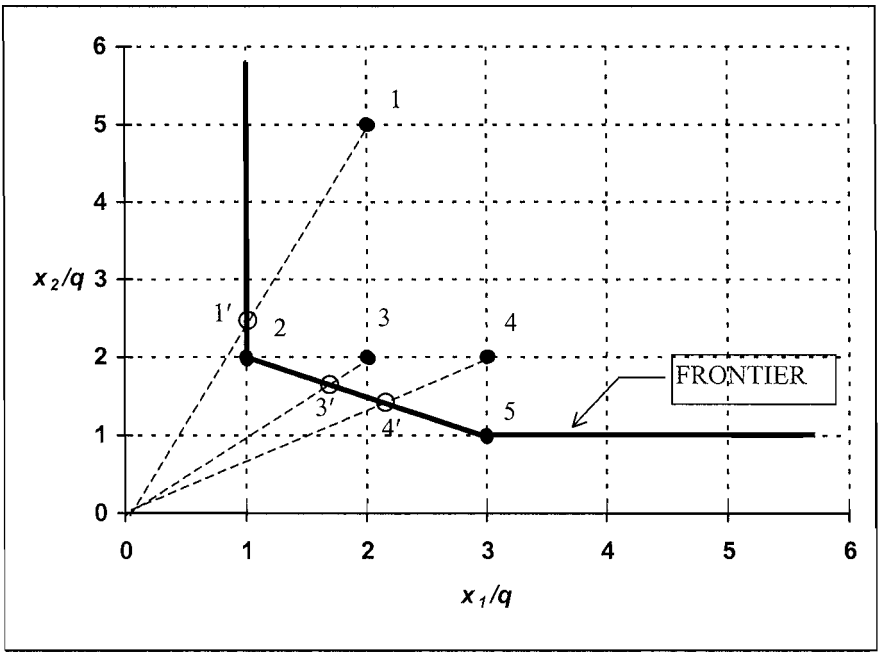

Figure 6.2 CRS Input-Orientated DEA Example

Table 6.2 CRS Input-Orientated DEA Results

firm	θ	λ_1	λ_2	λ_3	λ_4	λ_5	IS_1	IS_2	OS
1	0.5	-	0.5	-	-	-	-	0.5	-
2	1.0	-	1.0	-	-	-	-	-	-
3	0.833	-	1.0	-	-	0.5	-	-	-
4	0.714	-	0.214	-	-	0.286	-	-	-
5	1.0	-	-	-	-	1.0	-	-	-

Furthermore, note that in Table 6.2 firms 2 and 5 have TE values of 1.0 and that their peers are themselves. This is as one would expect for the efficient points that define the frontier.

DEA Calculations using the Computer

DEA calculations can be conducted using a number of different computer programs. If you are familiar with linear programming (LP), then all you need is access to computer software that can conduct LP. For example, one could use spreadsheet software such as Excel or statistical software such as SAS or SHAZAM. There are also a few specialist DEA computer packages available, such as ONFront, IDEAS,

Frontier Analyst, Warwick DEA and DEAP.[10] We use DEAP Version 2.1 in this book. A brief description of the DEAP computer program is provided in the Appendix.

The DEAP computer program has a similar structure to the TFPIP computer program introduced in Chapter 4. To calculate the above simple numerical example using DEAP, the user must construct a data file and an instruction file. Note that all data, instruction and output files are text files. The data file for this example, EG1-DTA.TXT, (refer to Table 6.3a) contains five observations on one output and two inputs. The output quantities are listed in the first column and the inputs in the next two columns. These data are identical to those listed in Table 6.1.

The instruction file, EG1-INS.TXT, is listed in Table 6.3b. The purpose of the majority of entries in the file should be self explanatory, due to the comments on the right-hand side of the file.[11] The first two lines of the file contain the name of the data file (EG1-DTA.TXT) and an output file name (here we have used EG1-OUT.TXT). Then on the next four lines we specify the number of firms (5); number of time periods (1);[12] number of outputs (1); and number of inputs (2). On the last three lines, we specify a "0" to indicate CRS; a "0" to indicate an input orientation; and a "0" to indicate that we wish to estimate a standard DEA model.[13]

Table 6.3a Listing of Data File, EG1-DTA.TXT

```
1 2 5
2 2 4
3 6 6
1 3 2
2 6 2
```

Finally, we execute DEAP and type in the name of the instruction file (EG1-INS.TXT). The program sends the output to the file (EG1-OUT.TXT). This file is reproduced in Table 6.3c. The information presented in this output file should be self explanatory given the preceding discussion. Note that the results are identical to those presented in Table 6.2.

[10] For a discussion of the relative merits of various DEA computer programs, including information on their associated web sites, see the systematic review provided in Hollingsworth (2004).
[11] The comments in this instruction file are not read by the program.
[12] Note that the number of time periods must be equal to 1 unless the Malmquist DEA option is selected.
[13] Note that by specifying "0" on the final line we are asking that slacks be calculated using the multi-stage method. If we wished the 1-stage or 2-stage methods to be used we would have used a "3" or a "4", respectively. These different methods for the calculation of slacks are discussed further in the following chapter.

Table 6.3b Listing of Instruction File, EG1-INS.TXT

```
eg1-dta.txt     DATA FILE NAME
eg1-out.txt     OUTPUT FILE NAME
5               NUMBER OF FIRMS
1               NUMBER OF TIME PERIODS
1               NUMBER OF OUTPUTS
2               NUMBER OF INPUTS
0               0=INPUT AND 1=OUTPUT ORIENTATED
0               0=CRS AND 1=VRS
0               0=DEA(MULTI-STAGE), 1=COST-DEA, 2=MALMQUIST-DEA,
                       3=DEA(1-STAGE), 4=DEA(2-STAGE)
```

Table 6.3c Listing of Output File, EG1-OUT.TXT

```
Results from DEAP Version 2.1

Instruction file = eg1-ins.txt
Data file        = eg1-dta.txt

Input orientated DEA

Scale assumption: CRS

Slacks calculated using multi-stage method


EFFICIENCY SUMMARY:

 firm     te
    1  0.500
    2  1.000
    3  0.833
    4  0.714
    5  1.000

 mean  0.810


SUMMARY OF OUTPUT SLACKS:

 firm  output:          1
    1               0.000
    2               0.000
    3               0.000
    4               0.000
    5               0.000

 mean               0.000
```

```
SUMMARY OF INPUT SLACKS:

 firm  input:            1           2
    1                0.000       0.500
    2                0.000       0.000
    3                0.000       0.000
    4                0.000       0.000
    5                0.000       0.000

 mean                0.000       0.100


SUMMARY OF PEERS:

 firm  peers:
    1      2
    2      2
    3      5    2
    4      5    2
    5      5


SUMMARY OF PEER WEIGHTS:
  (in same order as above)

 firm  peer weights:
    1   0.500
    2   1.000
    3   0.500 1.000
    4   0.286 0.214
    5   1.000


PEER COUNT SUMMARY:
  (i.e., no. times each firm is a peer for another)

 firm  peer count:
    1       0
    2       3
    3       0
    4       0
    5       2


SUMMARY OF OUTPUT TARGETS:

 firm  output:           1
    1                1.000
    2                2.000
    3                3.000
    4                1.000
    5                2.000


SUMMARY OF INPUT TARGETS:

 firm  input:            1           2
    1                1.000       2.000
    2                2.000       4.000
    3                5.000       5.000
    4                2.143       1.429
    5                6.000       2.000
```

```
FIRM BY FIRM RESULTS:

Results for firm:     1
Technical efficiency = 0.500
PROJECTION SUMMARY:
  variable          original        radial         slack     projected
                       value      movement      movement         value
 output   1           1.000         0.000         0.000         1.000
 input    1           2.000        -1.000         0.000         1.000
 input    2           5.000        -2.500        -0.500         2.000
LISTING OF PEERS:
 peer   lambda weight
    2      0.500

Results for firm:     2
Technical efficiency = 1.000
PROJECTION SUMMARY:
  variable          original        radial         slack     projected
                       value      movement      movement         value
 output   1           2.000         0.000         0.000         2.000
 input    1           2.000         0.000         0.000         2.000
 input    2           4.000         0.000         0.000         4.000
LISTING OF PEERS:
 peer   lambda weight
    2      1.000

Results for firm:     3
Technical efficiency = 0.833
PROJECTION SUMMARY:
  variable          original        radial         slack     projected
                       value      movement      movement         value
 output   1           3.000         0.000         0.000         3.000
 input    1           6.000        -1.000         0.000         5.000
 input    2           6.000        -1.000         0.000         5.000
LISTING OF PEERS:
 peer   lambda weight
    5      0.500
    2      1.000

Results for firm:     4
Technical efficiency = 0.714
PROJECTION SUMMARY:
  variable          original        radial         slack     projected
                       value      movement      movement         value
 output   1           1.000         0.000         0.000         1.000
 input    1           3.000        -0.857         0.000         2.143
 input    2           2.000        -0.571         0.000         1.429
LISTING OF PEERS:
 peer   lambda weight
    5      0.286
    2      0.214

Results for firm:     5
Technical efficiency = 1.000
PROJECTION SUMMARY:
  variable          original        radial         slack     projected
                       value      movement      movement         value
 output   1           2.000         0.000         0.000         2.000
 input    1           6.000         0.000         0.000         6.000
 input    2           2.000         0.000         0.000         2.000
LISTING OF PEERS:
 peer   lambda weight
    5      1.000
```

6.3 The Variable Returns to Scale Model and Scale Efficiencies

The CRS assumption is appropriate when all firms are operating at an optimal scale. However, imperfect competition, government regulations, constraints on finance, etc., may cause a firm to be not operating at optimal scale. Various authors, such as Afriat (1972), Färe, Grosskopf and Logan (1983) and Banker, Charnes and Cooper (1984) suggested adjusting the CRS DEA model to account for variable returns to scale (VRS) situations. The use of the CRS specification when not all firms are operating at the optimal scale, results in measures of TE that are confounded by *scale efficiencies* (SE). The use of the VRS specification permits the calculation of TE devoid of these SE effects.

The CRS linear programming problem can be easily modified to account for VRS by adding the convexity constraint: $\mathbf{I1}'\boldsymbol{\lambda}=1$ to equation 6.3 to provide:

$$\begin{aligned} &\min_{\theta,\lambda} \theta, \\ &\text{st} \quad -\mathbf{q}_i + \mathbf{Q}\boldsymbol{\lambda} \geq \mathbf{0}, \\ &\qquad \theta\mathbf{x}_i - \mathbf{X}\boldsymbol{\lambda} \geq \mathbf{0}, \\ &\qquad \mathbf{I1}'\boldsymbol{\lambda}=1 \\ &\qquad \boldsymbol{\lambda} \geq \mathbf{0}, \end{aligned} \tag{6.5}$$

where $\mathbf{I1}$ is an $I\times1$ vector of ones. This approach forms a convex hull of intersecting planes[14] that envelope the data points more tightly than the CRS conical hull and thus provides technical efficiency scores that are greater than or equal to those obtained using the CRS model.

Note that the convexity constraint ($\mathbf{I1}'\boldsymbol{\lambda}=1$) essentially ensures that an inefficient firm is only "benchmarked" against firms of a similar size. That is, the projected point (for that firm) on the DEA frontier is a *convex* combination of observed firms. This convexity restriction is not imposed in the CRS case. Hence, in a CRS DEA, a firm may be benchmarked against firms that are substantially larger (smaller) than it. In this instance, the λ-weights sum to a value less than (greater than) one.

Calculation of Scale Efficiencies

Scale efficiency measures can be obtained for each firm by conducting both a CRS and a VRS DEA, and then decomposing the TE scores obtained from the CRS DEA into two components, one due to scale inefficiency and one due to "pure" technical inefficiency (ie. VRS TE). If there is a difference in the CRS and VRS TE scores for a particular firm, then this indicates that the firm has scale inefficiency.

[14] The use of the term "planes" is correct in the three-dimensional case. However, when we have more dimensions (i.e. when the number of input plus output variables exceeds three), the term "facet" is more appropriate.

In Figure 6.3, we illustrate scale inefficiency calculations using a one-input, one-output example. The CRS and VRS DEA frontiers are indicated in the figure. Under CRS, the input-orientated technical inefficiency of the point P is the distance PP_C. However, under VRS, the technical inefficiency would only be PP_V. The difference between these two TE measures, P_CP_V, is due to scale inefficiency. These concepts can be expressed in ratio efficiency measures as:

$$\mathrm{TE}_{\mathrm{CRS}} = AP_C/AP$$

$$\mathrm{TE}_{\mathrm{VRS}} = AP_V/AP$$

$$\mathrm{SE} = AP_C/AP_V$$

where all of these measures are bounded by zero and one. We also note that

$$\mathrm{TE}_{\mathrm{CRS}} = \mathrm{TE}_{\mathrm{VRS}} \times \mathrm{SE}$$

because

$$AP_C/AP = (AP_V/AP) \times (AP_C/AP_V).$$

Thus, the CRS technical efficiency measure is decomposed into "pure" technical efficiency and scale efficiency. This scale efficiency measure can be roughly interpreted as the ratio of the average product of a firm operating at the point P_V to the average product of the point operating at a point of (technically) optimal scale (point R).

The Nature of Returns to Scale

One shortcoming of this measure of scale efficiency is that the value does not indicate whether the firm is operating in an area of increasing or decreasing returns to scale. This latter issue can be determined by running an additional DEA problem with non-increasing returns to scale (NIRS) imposed. This is done by altering the DEA model in LP 6.5 by substituting the $\mathbf{I1}'\boldsymbol{\lambda} = 1$ restriction with $\mathbf{I1}'\boldsymbol{\lambda} \leq 1$, to provide:

$$\begin{aligned} &\min_{\theta,\lambda} \theta, \\ &\text{st} \quad -\mathbf{q}_i + \mathbf{Q}\boldsymbol{\lambda} \geq \mathbf{0}, \\ &\qquad \theta\mathbf{x}_i - \mathbf{X}\boldsymbol{\lambda} \geq \mathbf{0}, \\ &\qquad \mathbf{I1}'\boldsymbol{\lambda} \leq 1 \\ &\qquad \boldsymbol{\lambda} \geq \mathbf{0}, \end{aligned} \tag{6.6}$$

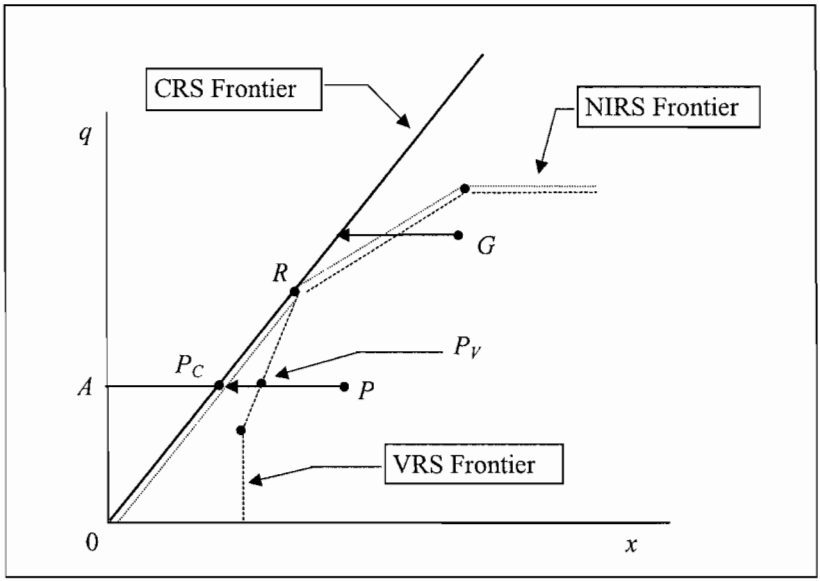

Figure 6.3: Scale Efficiency Measurement in DEA

The NIRS DEA frontier is also plotted in Figure 6.3. The nature of the scale inefficiencies (i.e., due to increasing or decreasing returns to scale) for a particular firm can be determined by seeing whether the NIRS TE score is equal to the VRS TE score. If they are unequal (as is the case for the point *P* in Figure 6.3) then increasing returns to scale exist for that firm. If they are equal (as is the case for point *G* in Figure 6.3) then decreasing returns to scale apply.[15] Examples of this approach, applied to the electricity industry, are provided in Färe, Grosskopf and Logan (1983, 1985).

Note that the constraint, $\mathbf{I1}'\boldsymbol{\lambda} \leq 1$, ensures that the *i*-th firm is not "benchmarked" against firms that are substantially larger than it, but may be compared with firms smaller than it.

Example 2

This is a simple numerical example involving five firms that produce a single output using a single input. The data are listed in Table 6.4 and the VRS and CRS input-orientated DEA results are listed in Table 6.5 and plotted in Figure 6.4. Given that we are using an input orientation, the efficiencies are measured horizontally across Figure 6.4. We observe that firm 3 is the only efficient firm (i.e., on the DEA frontier) when CRS is assumed, but that firms 1, 3 and 5 are efficient when VRS is assumed.

[15] Note that if $TE_{CRS}=TE_{VRS}$, then, by definition, the firm is operating under CRS.

Table 6.4 Example Data for VRS DEA

Firm	q	x
1	1	2
2	2	4
3	3	3
4	4	5
5	5	6

The calculation of the various efficiency measures can be illustrated using firm 2, which is inefficient under both CRS and VRS technologies. The CRS technical efficiency (TE) is equal to 2/4=0.5; the VRS TE is 2.5/4=0.625 and the scale efficiency is equal to the ratio of the CRS TE to the VRS TE, which is 0.5/0.625=0.8. We also observe that firm 2 is on the increasing returns to scale (IRS) portion of the VRS frontier.

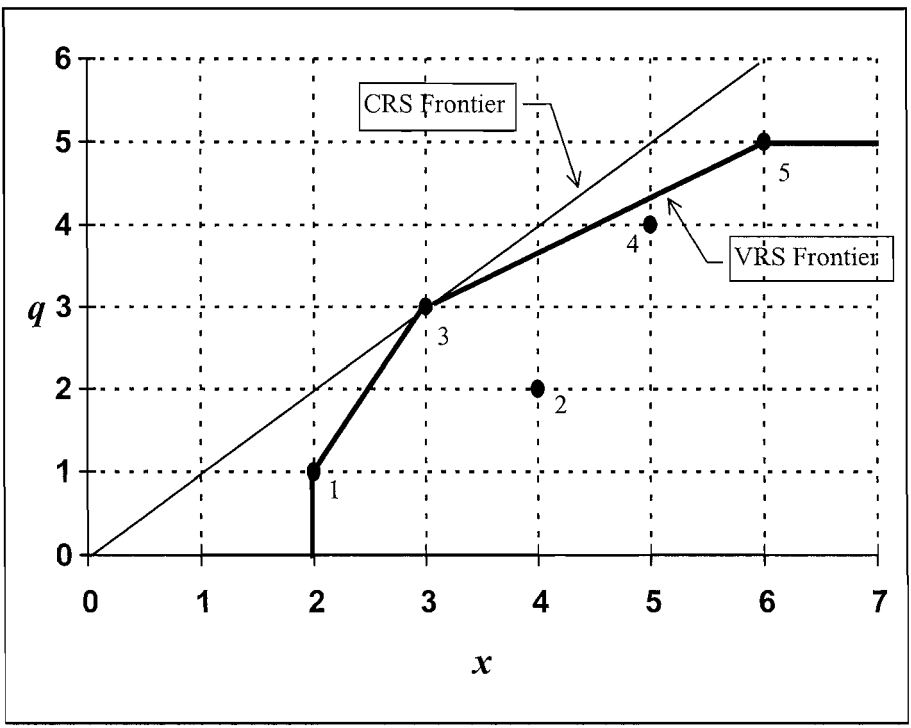

Figure 6.4 VRS Input-Orientated DEA Example

Table 6.5 VRS Input-Orientated DEA Results

Firm	CRS TE	VRS TE	Scale	
1	0.500	1.000	0.500	IRS
2	0.500	0.625	0.800	IRS
3	1.000	1.000	1.000	-
4	0.800	0.900	0.889	DRS
5	0.833	1.000	0.833	DRS
mean	**0.727**	**0.905**	**0.804**	

We now illustrate these calculations using the DEAP computer program. The data file for this example, EG2-DTA.TXT (refer to Table 6.6a), contains five observations on one output and one input. The output quantities are listed in the first column and the inputs in the second column. These data are identical to those listed in Table 6.4.

The DEAP instruction file, EG2-INS.TXT, is listed in Table 6.6b. The only changes relative to EG1-INS.TXT are that:

- the input and output file names are different;
- the number of inputs is reduced to 1; and
- there is a "1" entered on the second last line to indicate that VRS is required.

The output file, EG1-OUT.TXT, is reproduced in Table 6.6c. These results are identical to those presented in Table 6.5. Note that when the VRS option is specified, the DEAP program conducts VRS, CRS and NIRS DEA and calculates scale efficiencies as well as technical efficiencies.

Table 6.6a Listing of Data File, EG2-DTA.TXT

```
1 2
2 4
3 3
4 5
5 6
```

Table 6.6b Listing of Instruction File, EG2-INS.TXT

```
eg2-dta.txt     DATA FILE NAME
eg2-out.txt     OUTPUT FILE NAME
5               NUMBER OF FIRMS
1               NUMBER OF TIME PERIODS
1               NUMBER OF OUTPUTS
1               NUMBER OF INPUTS
0               0=INPUT AND 1=OUTPUT ORIENTATED
1               0=CRS AND 1=VRS
0               0=DEA(MULTI-STAGE), 1=COST-DEA, 2=MALMQUIST-DEA,
                       3=DEA(1-STAGE), 4=DEA(2-STAGE)
```

Table 6.6c Listing of Output File, EG2-OUT.TXT

```
Results from DEAP Version 2.1

Instruction file = eg2-ins.txt
Data file        = eg2-dta.txt

Input orientated DEA

Scale assumption: VRS

Slacks calculated using multi-stage method

EFFICIENCY SUMMARY:

 firm  crste  vrste  scale

    1  0.500  1.000  0.500 irs
    2  0.500  0.625  0.800 irs
    3  1.000  1.000  1.000  -
    4  0.800  0.900  0.889 drs
    5  0.833  1.000  0.833 drs

 mean  0.727  0.905  0.804

Note: crste = technical efficiency from CRS DEA
      vrste = technical efficiency from VRS DEA
      scale = scale efficiency = crste/vrste

Note also that all subsequent tables refer to VRS results


SUMMARY OF OUTPUT SLACKS:

 firm  output:          1
    1               0.000
    2               0.000
    3               0.000
    4               0.000
    5               0.000

 mean               0.000
```

```
SUMMARY OF INPUT SLACKS:

 firm  input:            1
    1                0.000
    2                0.000
    3                0.000
    4                0.000
    5                0.000

 mean                0.000


SUMMARY OF PEERS:

 firm  peers:
    1      1
    2      1    3
    3      3
    4      3    5
    5      5


SUMMARY OF PEER WEIGHTS:
  (in same order as above)

 firm  peer weights:
    1   1.000
    2   0.500 0.500
    3   1.000
    4   0.500 0.500
    5   1.000


PEER COUNT SUMMARY:
  (i.e., no. times each firm is a peer for another)

 firm  peer count:
    1       1
    2       0
    3       2
    4       0
    5       1


SUMMARY OF OUTPUT TARGETS:

 firm  output:           1
    1                1.000
    2                2.000
    3                3.000
    4                4.000
    5                5.000


SUMMARY OF INPUT TARGETS:

 firm  input:            1
    1                2.000
    2                2.500
    3                3.000
    4                4.500
    5                6.000
```

```
FIRM BY FIRM RESULTS:

Results for firm:      1
Technical efficiency = 1.000
Scale efficiency     = 0.500  (irs)
PROJECTION SUMMARY:
  variable          original        radial         slack     projected
                       value      movement      movement         value
 output   1            1.000         0.000         0.000         1.000
 input    1            2.000         0.000         0.000         2.000
LISTING OF PEERS:
 peer   lambda weight
    1      1.000

Results for firm:      2
Technical efficiency = 0.625
Scale efficiency     = 0.800  (irs)
PROJECTION SUMMARY:
  variable          original        radial         slack     projected
                       value      movement      movement         value
 output   1            2.000         0.000         0.000         2.000
 input    1            4.000        -1.500         0.000         2.500
LISTING OF PEERS:
 peer   lambda weight
    1      0.500
    3      0.500

Results for firm:      3
Technical efficiency = 1.000
Scale efficiency     = 1.000  (crs)
PROJECTION SUMMARY:
  variable          original        radial         slack     projected
                       value      movement      movement         value
 output   1            3.000         0.000         0.000         3.000
 input    1            3.000         0.000         0.000         3.000
LISTING OF PEERS:
 peer   lambda weight
    3      1.000

Results for firm:      4
Technical efficiency = 0.900
Scale efficiency     = 0.889  (drs)
PROJECTION SUMMARY:
  variable          original        radial         slack     projected
                       value      movement      movement         value
 output   1            4.000         0.000         0.000         4.000
 input    1            5.000        -0.500         0.000         4.500
LISTING OF PEERS:
 peer   lambda weight
    3      0.500
    5      0.500

Results for firm:      5
Technical efficiency = 1.000
Scale efficiency     = 0.833  (drs)
PROJECTION SUMMARY:
  variable          original        radial         slack     projected
                       value      movement      movement         value
 output   1            5.000         0.000         0.000         5.000
 input    1            6.000         0.000         0.000         6.000
LISTING OF PEERS:
 peer   lambda weight
    5      1.000
```

6.4 Input and Output Orientations

In the preceding input-orientated models, discussed in Sections 6.2 and 6.3, the method sought to identify technical inefficiency as a proportional reduction in input usage, with output levels held constant. This corresponds to Farrell's input-based measure of technical inefficiency. As discussed in Chapter 3, it is also possible to measure technical inefficiency as a proportional increase in output production, with input levels held fixed. The two measures provide the same value under CRS but are unequal when VRS is assumed. Given that linear programming does not suffer from such statistical problems as simultaneous-equations bias, the choice of an appropriate orientation is not as crucial as it is in the case of econometric estimation. In a number of studies, analysts have tended to select input-orientated models because many firms have particular orders to fill (e.g., as in electricity generation) and, hence, the input quantities appear to be the primary decision variables, although this argument may not be as strong in all industries. In some industries, the firms may be given a fixed quantity of resources and asked to produce as much output as possible. In this case, an output orientation would be more appropriate. Essentially, one should select the orientation according to which quantities (inputs or outputs) the managers have most control over. Furthermore, in many instances, the choice of orientation has only a minor influence upon the scores obtained (e.g., see Coelli and Perelman, 1999).

The output-orientated DEA models are very similar to their input-orientated counterparts. Consider the example of the following output-orientated VRS model:

$$\begin{aligned} \max_{\phi,\lambda}\ & \phi, \\ \text{st}\quad & -\phi\mathbf{q}_i + \mathbf{Q}\boldsymbol{\lambda} \geq 0, \\ & \mathbf{x}_i - \mathbf{X}\boldsymbol{\lambda} \geq 0, \\ & \mathbf{I1}'\boldsymbol{\lambda}=1 \\ & \boldsymbol{\lambda} \geq \mathbf{0}, \end{aligned} \tag{6.7}$$

where $1 \leq \phi < \infty$, and ϕ-1 is the proportional increase in outputs that could be achieved by the i-th firm, with input quantities held constant.[16] Note that $1/\phi$ defines a TE score that varies between zero and one (and that this is the output-orientated TE score reported by DEAP).

A two-output example of an output-orientated DEA could be represented by a piece-wise linear production possibility curve, such as that depicted in Figure 6.5. Note that the observations lie *below* this curve, and that the sections of the curve that are at right angles to the axes result in output slack being calculated when a production point is projected onto those parts of the curve by a radial expansion in outputs. For example, the point P is projected to the point P', which is on the frontier but not on the *efficient frontier*. This is because the production of q_1 could

[16] An output-orientated CRS model is defined in a similar way, but it is not presented here for brevity.

be increased by the amount AP' without using any more inputs. Thus there is output slack in this case of AP' in output q_1.

One point of importance is that *the output- and input-orientated DEA models will estimate exactly the same frontier and therefore, by definition, identify the same set of firms as being efficient. It is only the efficiency measures associated with the inefficient firms that may differ between the two methods.* The two types of measures were described in Chapter 3, where we observe that the two measures provide equivalent values only under constant returns to scale.

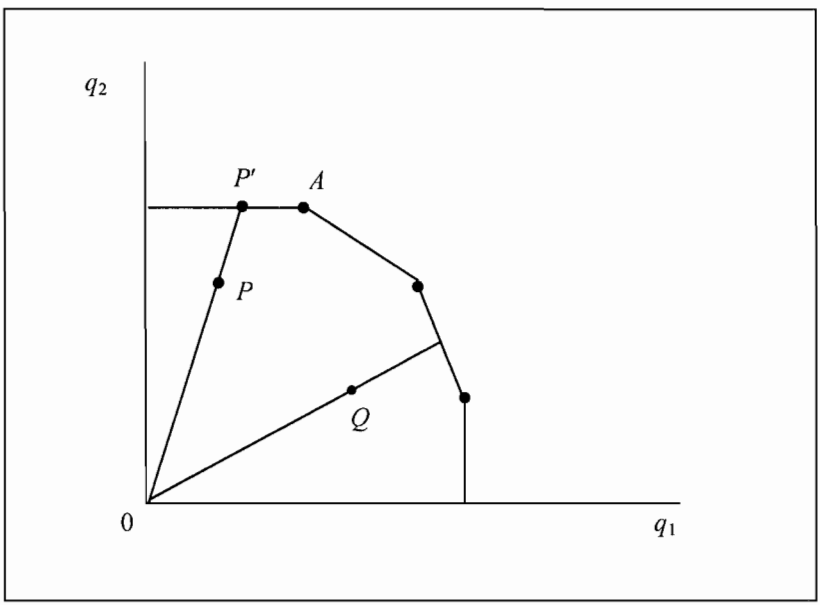

Figure 6.5 Output-Orientated DEA

6.5 Conclusions

In this chapter, we provide a brief introduction to the basic DEA models. Namely, the input- and output-orientated CRS and VRS models. We discuss how these models can be used to measure technical and scale efficiencies and how one can use NIRS DEA to help identify the nature of scale economies. Terminologies, such as peers, targets and slacks, are also introduced.

In the following chapter, we discuss some of the ways in which these basic DEA models can be extended. We discuss allocative efficiency, environmental variables, non-discretionary variables, the treatment of slacks, super-efficiency measures, and a variety of additional issues.

7. ADDITIONAL TOPICS ON DATA ENVELOPMENT ANALYSIS

7.1 Introduction

This chapter continues the discussion of data envelopment analysis (DEA) that began in the previous chapter. In that chapter, we described how one could calculate various efficiency measures using the linear programming approach known as DEA. We discussed the basic constant returns to scale (CRS) and variable returns to scale (VRS) DEA models from both the input- and output-orientations.

In this chapter we discuss some popular extensions of these basic DEA models. The extensions we consider involve allocative efficiency, non-discretionary variables, environmental variables, the treatment of slacks and congestion efficiency, weights restrictions, super efficiency and bootstrap methods. We also provide an empirical example using data on Australian universities.

7.2 Price Information and Allocative Efficiency

If price data are available and a behavioural objective, such as cost minimisation or revenue or profit maximisation, is appropriate, then it is possible to measure allocative efficiencies as well as technical efficiencies. To achieve this, two sets of linear programs are required, one to measure technical efficiency and the other to measure economic efficiency. The cases of cost minimisation, revenue maximisation and profit maximisation are now discussed.

7.2.1 Cost Minimisation

For the case of VRS cost minimisation, the input-orientated DEA model, defined in LP 6.5, is conducted to obtain technical efficiencies (TE). The next step requires the solution of the following cost minimisation DEA:

$$\begin{aligned} \min_{\lambda, \mathbf{x}_i^*} \quad & \mathbf{w}_i'\mathbf{x}_i^*, \\ \text{st} \quad & -\mathbf{q}_i + \mathbf{Q}\boldsymbol{\lambda} \geq \mathbf{0}, \\ & \mathbf{x}_i^* - \mathbf{X}\boldsymbol{\lambda} \geq \mathbf{0}, \\ & \mathbf{I}1'\boldsymbol{\lambda}=1 \\ & \boldsymbol{\lambda} \geq \mathbf{0}, \end{aligned} \tag{7.1}$$

where $\mathbf{w}_i$ is a $N\times1$ vector of input prices for the i-th firm and $\mathbf{x}_i^*$ (which is calculated by the LP) is the cost-minimising vector of input quantities for the i-th firm, given the input prices $\mathbf{w}_i$ and the output levels $\mathbf{q}_i$, and all other notation is as previously defined in Chapter 6.

The total cost efficiency (CE) of the i-th firm is calculated as

$$\text{CE} = \mathbf{w}_i'\mathbf{x}_i^* / \mathbf{w}_i'\mathbf{x}_i.$$

That is, CE is the ratio of minimum cost to observed cost, for the i-th firm.

The (input-mix) allocative efficiency is then calculated residually as

$$\text{AE} = \text{CE/TE}.$$

These three measures (TE, AE and CE) can take values ranging from 0 to 1, where a value of 1 indicates full efficiency.

Note that this procedure implicitly includes any slacks into the allocative efficiency measure. This is often justified on the grounds that the slacks reflect inappropriate input mixes (see Ferrier and Lovell, 1990, p.235).

7.2.2 Revenue Maximisation

If revenue maximisation is a more appropriate behavioural assumption, then allocative inefficiency in output-mix selection can be accounted for in a similar manner. For the case of VRS revenue maximisation, technical efficiencies are calculated by solving the output-orientated DEA model, defined in LP 6.7. The following revenue maximisation DEA problem is then solved:

$$\begin{aligned} \max_{\lambda, yi^*} \quad & \mathbf{p}_i'\mathbf{q}_i^*, \\ \text{st} \quad & -\mathbf{q}_i^* + \mathbf{Q}\boldsymbol{\lambda} \geq \mathbf{0}, \\ & \mathbf{x}_i - \mathbf{X}\boldsymbol{\lambda} \geq \mathbf{0}, \\ & \mathbf{I1}'\boldsymbol{\lambda}=1, \\ & \boldsymbol{\lambda} \geq \mathbf{0}, \end{aligned} \tag{7.2}$$

where $\mathbf{p}_i$ is a $M\times 1$ vector of input prices for the i-th firm and $\mathbf{q}_i^*$ (which is calculated by the LP) is the revenue-maximising vector of output quantities for the i-th firm, given the output prices $\mathbf{p}_i$ and the input levels $\mathbf{x}_i$. The total revenue efficiency (RE) of the i-th firm is calculated as

$$\text{RE} = \mathbf{p}_i'\mathbf{q}_i / \mathbf{p}_i'\mathbf{q}_i^*$$

That is, RE is the ratio of observed revenue to maximum revenue.

The (output-mix) allocative efficiency measure is then obtained residually as

$$\text{AE} = \text{RE/TE}.$$

These three measures (TE, AE and RE) can take values ranging from 0 to 1, where a value of 1 indicates full efficiency.

7.2.3 Profit Maximisation

If one has access to price data on both inputs and outputs, then one can also calculate profit efficiency using DEA methods. The profit maximisation DEA problem is specified as follows:

$$\begin{aligned} \max_{\lambda, yi^*, xi^*} \quad & (\mathbf{p}_i'\mathbf{q}_i^* - \mathbf{w}_i'\mathbf{x}_i^*) \\ \text{st} \quad & -\mathbf{q}_i^* + \mathbf{Q}\boldsymbol{\lambda} \geq \mathbf{0}, \\ & \mathbf{x}_i^* - \mathbf{X}\boldsymbol{\lambda} \geq \mathbf{0}, \\ & \mathbf{I1}'\boldsymbol{\lambda}=1, \\ & \boldsymbol{\lambda} \geq \mathbf{0}, \end{aligned} \tag{7.3}$$

where all notation used is as previously defined.

Once the profit maximizing point for each firm is identified ($\mathbf{q}_i^*$, $\mathbf{x}_i^*$), one could then specify a profit efficiency measure as the ratio of observed profit over maximum (potential) profit. However, this measure need not be bounded by 0 and 1. It could be negative if profits are negative, or it could be undefined if maximum profit is 0.

Furthermore, the decomposition of a profit efficiency measure into technical and allocative components is not straight forward either. There are a number of possible

choices. For example, TE could be measured with either an input or output orientation. However, this choice will affect the decomposition results.

Färe, Grosskopf and Weber (2004) address this issue by making use of *directional distance functions*. Their approach requires the solution of two sets of linear programs. The first involves a profit maximising DEA to measure profit efficiency and the second DEA is one in which technical efficiency is measured as a simultaneous reduction in the input vector and expansion of the output vector, using the directional distance function construct. This approach does help one avoid making an arbitrary choice between input and output orientated technical efficiency measures, but the use of directional distance functions requires one to specify appropriate "directions" for each firm, which can be a source of some debate, and hence can affect the decomposition obtained.

7.2.4 A CRS Cost Efficiency DEA Example

In this example, we take the data from the two-input, one-output, input-orientated DEA example in Table 6.1 and add the information that all firms face the same prices, which are 1 and 3 for inputs 1 and 2, respectively. The solution of this CRS cost efficiency DEA problem is illustrated in Figure 7.1. This figure is equivalent to Figure 6.2 except that the isocost line with a slope of -1/3 is also drawn so that it is tangential to the isoquant. From this diagram, we observe that firm 5 is the only cost efficient firm and that all other firms have some allocative inefficiency. The various cost efficiencies and allocative efficiencies are listed in Table 7.1.

The calculation of these efficiencies can be illustrated using firm 3. We noted in Chapter 6 that the technical efficiency of firm 3 is measured along the ray from the origin (0) to the point 3 and that it is equal to the ratio of the distance from 0 to the point 3′ over the distance from 0 to the point 3 and that this is equal to 0.833. The allocative efficiency is equal to the ratio of the distances 0 to 3″ over 0 to 3′, which is equal to 0.9. The cost efficiency is the ratio of distances 0 to 3″ over 0 to 3, which is equal to 0.75. We also note that $0.833 \times 0.9 = 0.750$.

The DEAP instructions needed to calculate the technical, allocative and cost efficiency measures, reported in Table 7.1, are minor variants of those presented in the previous chapter. The data file, EG3-DTA.TXT (refer to Table 7.2a), contains the same data as EG1-DTA.TXT plus two additional columns. Recall that there are five observations on one output and two inputs. The output quantities are listed in the first column and the input quantities in the next two columns. In addition to this, two columns containing input price data are listed to the right of these. In this particular example we assume all firms face the same prices and that these prices are 1 and 3 for inputs 1 and 2, respectively.

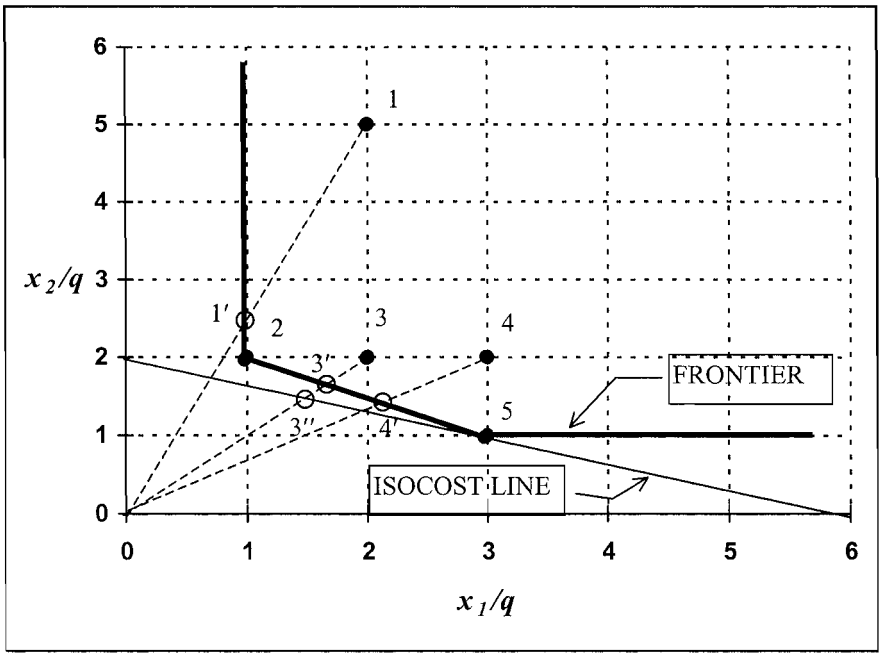

Figure 7.1 CRS Cost Efficiency DEA Example

Table 7.1 CRS Cost Efficiency DEA Results

firm	technical efficiency	allocative efficiency	cost efficiency
1	0.500	0.706	0.353
2	1.000	0.857	0.857
3	0.833	0.900	0.750
4	0.714	0.933	0.667
5	1.000	1.000	1.000
mean	**0.810**	**0.879**	**0.725**

The EG3-INS.TXT file is listed in Table 7.2b. The only changes in this instruction file relative to the EG1-INS.TXT file, presented in Table 6.3b, are that:

- the input and output file names are different;
- there is a "1" entered on the last line to indicate that a cost efficiency DEA is required.

The output file, EG3-OUT.TXT, is reproduced in Table 7.2c. The technical efficiency results are identical to those listed in EG1-OUT.TXT in Table 6.3c.

However, the allocative and cost efficiencies are now also listed. Furthermore, a table of input quantities for minimum cost production are now also listed. Note that all of the optimal input quantities are used in the same ratio (3:1) because each firm faces the same relative input prices. If the prices differ for different firms then the optimal input ratios are then likely to be different.

Table 7.2a Listing of Data File, EG3-DTA.TXT

```
1 2 5 1 3
2 2 4 1 3
3 6 6 1 3
1 3 2 1 3
2 6 2 1 3
```

Table 7.2b Listing of Instruction File, EG3-INS.TXT

```
eg3-dta.txt     DATA FILE NAME
eg3-out.txt     OUTPUT FILE NAME
5               NUMBER OF FIRMS
1               NUMBER OF TIME PERIODS
1               NUMBER OF OUTPUTS
2               NUMBER OF INPUTS
0               0=INPUT AND 1=OUTPUT ORIENTATED
0               0=CRS AND 1=VRS
1               0=DEA(MULTI-STAGE), 1=COST-DEA, 2=MALMQUIST-DEA,
                       3=DEA(1-STAGE), 4=DEA(2-STAGE)
```

7.3 Non-Discretionary Variables

In this section, we consider some alternative DEA models in which we specify that some variables are under the control of the manager and some are not. Consider the case of the VRS input orientated DEA model in LP 6.5. In that model, we assume that the manager faces fixed output quantities and variable input quantities. For example, this could be applicable to the case of dairy farming where the farmers face fixed production quotas (at least in the short run) but can vary their input quantities. However, in this dairy-farming example, we may also argue that the farmers have little control over certain input variables in the short run (e.g. land and buildings). We ask the question: By how much can the manager reduce variable input usage (eg. labour and other purchased materials and services) given a fixed level of output, land and buildings?

Table 7.2c Listing of Output File, EG3-OUT.TXT

```
Results from DEAP Version 2.1

Instruction file = eg3-ins.txt
Data file        = eg3-dta.txt

Cost efficiency DEA

Scale assumption: CRS


EFFICIENCY SUMMARY:

 firm     te     ae     ce
    1  0.500  0.706  0.353
    2  1.000  0.857  0.857
    3  0.833  0.900  0.750
    4  0.714  0.933  0.667
    5  1.000  1.000  1.000

 mean  0.810  0.879  0.725

Note: te = technical efficiency
      ae = allocative efficiency = ce/te
      ce = cost efficiency


SUMMARY OF COST MINIMISING INPUT QUANTITIES:

 firm  input:          1          2
    1               3.000      1.000
    2               6.000      2.000
    3               9.000      3.000
    4               3.000      1.000
    5               6.000      2.000
```

We formulate a DEA model in which we only seek radial reduction in the inputs over which the manager has discretionary control. In this case, we divide the inputs into discretionary and non-discretionary sets (denoted by $\mathbf{X}^D$ and $\mathbf{X}^{ND}$, respectively) and rewrite equation 6.5 (for the VRS case) as:

$$
\begin{aligned}
&\min_{\theta,\lambda} \theta, \\
&\text{st} \quad -\mathbf{q}_i + \mathbf{Q}\boldsymbol{\lambda} \geq \mathbf{0}, \\
&\qquad \theta\mathbf{x}_i^D - \mathbf{X}^D\boldsymbol{\lambda} \geq \mathbf{0}, \\
&\qquad \mathbf{x}_i^{ND} - \mathbf{X}^{ND}\boldsymbol{\lambda} \geq \mathbf{0}, \\
&\qquad \mathbf{I1}'\boldsymbol{\lambda}=1 \\
&\qquad \boldsymbol{\lambda} \geq 0,
\end{aligned}
\tag{7.4}
$$

In the above DEA problem, the θ-parameter is only associated with the discretionary inputs and, hence, the problem only seeks radial reduction in this subset of the inputs. This approach may be visualised in the two-input case by looking at Figure 6.1, and assuming that the capital input is on the vertical axis and labour is on the horizontal axis. If our discretionary set involved labour and our non-discretionary set involved capital, then the linear program in 7.4 would seek a reduction in the labour input only. For example, in the case of the firm operating at the point *B* in Figure 6.5, it would seek contraction in a horizontal direction (to the left) until the isoquant was reached. This would reduce the quantity of labour used, while holding the capital quantity fixed. For more on this approach, see Banker and Morey (1986a).

One can also apply this concept to the models involving price information described in the previous section. For example, one could measure the short run cost efficiency of firms by adjusting the linear program in 7.1 so that the levels of the non-discretionary inputs (such as capital) were held fixed. Using the notation defined above we specify:

$$
\begin{aligned}
&\min_{\lambda, x_i^{D*}} \ \mathbf{w}_i^{D\prime}\mathbf{x}_i^{D*}, \\
&\text{st} \quad -\mathbf{q}_i + \mathbf{Q}\boldsymbol{\lambda} \geq \mathbf{0}, \\
&\qquad \theta\mathbf{x}_i^{D*} - \mathbf{X}^{D}\boldsymbol{\lambda} \geq \mathbf{0}, \\
&\qquad \mathbf{x}_i^{ND} - \mathbf{X}^{ND}\boldsymbol{\lambda} \geq \mathbf{0}, \\
&\qquad \mathbf{I1}'\boldsymbol{\lambda}=1, \\
&\qquad \boldsymbol{\lambda} \geq 0.
\end{aligned}
\tag{7.5}
$$

The above discussion considers non-discretionary input variables. The approach can also be extended to the case of non-discretionary output variables in a similar manner.

7.4 Adjusting for the Environment

Here we use the term *environment* to describe factors that could influence the efficiency of a firm, where such factors are not traditional inputs and are assumed not under the control of the manager. Some examples of environmental variables include (see Fried, Schmidt and Yaisawarng, 1999):

1. ownership differences, such as public/private or corporate/non-corporate;
2. location characteristics, such as:
    - coal-fired electric power stations influenced by coal quality;
    - electric power distribution networks influenced by population density and average customer size;

- schools influenced by socio-economic status of children and city/country location; etc.

3. labour union power; and
4. government regulations.

There are a number of ways in which environmental variables can be accommodated in a DEA analysis. Some possible methods are discussed below.

Method 1

If the values of the environmental variable can be ordered from the least to the most detrimental effect upon efficiency, then the approach of Banker and Morey (1986a) can be followed. In this approach the efficiency of the i-th firm is compared with those firms in the sample that have a value of the environmental variable which is less than or equal to that of the i-th firm. For example, consider the case where an analyst is studying hamburger restaurants and the analyst believes that the type of location has an influence upon production. The analyst has information on whether the restaurant is located in a city centre, the suburbs or in a country area, and believes that the city is the most favourable location and that the country is the least favourable. In this instance, the analyst would restrict the comparison set to be: (i) only country restaurants for a country restaurant; (ii) only country and suburban restaurants for a suburban restaurant; and (iii) all restaurants for a city restaurant. This would ensure that no restaurant is compared with another restaurant that has a more favourable environment.

Method 2

If there is no natural ordering of the environmental variable (e.g., public versus private ownership) then one can use a method proposed by Charnes, Cooper and Rhodes (1981). This method involves three stages:

1. divide the sample into public/private sub-samples and solve DEAs for each sub-sample;
2. project all observed data points onto their respective frontiers; and
3. solve a single DEA using the projected points and assess any difference in the mean efficiency of the two sub-samples.

Note that one problem with Methods 1 and 2 is that the comparison set can be greatly reduced, resulting in many firms being found to be efficient and thus reducing the discriminating power of the analysis. Another problem is that only one environmental variable can be considered by these two methods. Method 2 has the additional problem that it requires that the environmental variable be a categorical variable, while Method 1 suffers from the problem that it requires that the direction of the influence of the environmental variable (upon efficiency) be known *a priori*.

In fact, in many instances the direction of influence is not known. For example, in an analysis of electricity utilities, we may be primarily interested in *determining* the influence of ownership status (public versus private) upon efficiency, and hence not wish to impose any a priori judgement.

Method 3

Another possible method is to include the environmental variable(s) directly into the LP formulation.[1] In general, an environmental variable is either:

i. included as a non-discretionary input or output variable, using the methods described in section 7.3, or alternatively
ii. included as a non-discretionary neutral variable.

If option (i) is used, we must first decide upon the direction of influence of the environmental variable. That is, are higher values of the variable likely to aide or impair efficiency? For example, in the case of electricity distribution, the density of the network (e.g. the population density in the city being served) is likely to have a favourable effect upon efficiency because a shorter network is needed to serve a particular number of customers. On the other hand, the number of thunderstorms per year is likely to have a detrimental effect upon efficiency because extra repair crews would be required.

If the variable is believed to have a positive effect upon efficiency then it should be included in the linear program in the same way as a non-discretionary input would be included. For example, consider the VRS input-orientated DEA problem in LP 6.5, and assume we have L "positive-effect" environmental variables to add to the model, and these are denoted by the $L\times 1$ vector $\mathbf{z}_i$ for the i-th firm and by the $L\times I$ matrix $\mathbf{Z}$ for the full sample. The input-orientated VRS DEA in LP 6.5 becomes:

$$\begin{aligned} \min_{\theta,\lambda}\ & \theta, \\ \text{st}\quad & -\mathbf{q}_i + \mathbf{Q}\boldsymbol{\lambda} \geq \mathbf{0}, \\ & \theta\mathbf{x}_i - \mathbf{X}\boldsymbol{\lambda} \geq \mathbf{0}, \\ & \mathbf{z}_i - \mathbf{Z}\boldsymbol{\lambda} \geq \mathbf{0}, \\ & \mathbf{I1}'\boldsymbol{\lambda}=1, \\ & \boldsymbol{\lambda} \geq \mathbf{0}. \end{aligned} \tag{7.6}$$

In this way, the i-th firm is compared with a theoretical firm that has an environment that is no better than that of the i-th firm.

On the other hand, if instead we have a set of "negative-effect" environmental variables to add to the model then they should be included in the linear program in the same way as a non-discretionary *output* would be included:

[1] For example, see Bessent and Bessent (1980) and Ferrier and Lovell (1990).

$$
\begin{aligned}
&\min_{\theta,\lambda} \theta, \\
&\text{st} \quad -\mathbf{q}_i + \mathbf{Q}\boldsymbol{\lambda} \geq \mathbf{0}, \\
&\qquad \theta\mathbf{x}_i - \mathbf{X}\boldsymbol{\lambda} \geq \mathbf{0}, \\
&\qquad -\mathbf{z}_i + \mathbf{Z}\boldsymbol{\lambda} \geq \mathbf{0}, \\
&\qquad \mathbf{I1}'\boldsymbol{\lambda}=1 \\
&\qquad \boldsymbol{\lambda} \geq \mathbf{0}.
\end{aligned} \tag{7.7}
$$

Of course, if we have a mixture of "positive effect and "negative effect" environmental variables then the linear program would be a mixture of these two models. Furthermore, this approach can be similarly applied in the case of output-orientated models.

Now we consider option (ii). If one is unsure as to the direction of the influence of the environmental variables, then the variables can be included in the LP problem in an equality form. For example:

$$
\begin{aligned}
&\min_{\theta,\lambda} \theta, \\
&\text{st} \quad -\mathbf{q}_i + \mathbf{Q}\boldsymbol{\lambda} \geq \mathbf{0}, \\
&\qquad \theta\mathbf{x}_i - \mathbf{X}\boldsymbol{\lambda} \geq \mathbf{0}, \\
&\qquad -\mathbf{z}_i + \mathbf{Z}\boldsymbol{\lambda} = \mathbf{0}, \\
&\qquad \mathbf{I1}'\boldsymbol{\lambda}=1 \\
&\qquad \boldsymbol{\lambda} \geq \mathbf{0}.
\end{aligned} \tag{7.8}
$$

This formulation ensures that the *i*-th firm is only compared with a (theoretical) frontier firm that has the same environment (no better and no worse). This approach avoids the necessity for one to pre-specify the direction of the influence of the environmental variable, which may be uncertain. However, it has the disadvantage that equality restrictions of this form can greatly reduce the reference set for each firm and hence inflate the efficiency scores obtained.

One can deal with this issue by noting that the signs on the dual variables[2] associated with the Z-variables indicate whether the variables have favourable or unfavourable effects upon production. Thus, one could first estimate the DEA model in LP 7.8 and then after observing the signs on the dual variables, decide whether the variables are a "positive-effect" or "negative-effect" variables and then re-run the model with the appropriate inequalities specified. Another possible solution to this problem is described as Method 4 below.

Finally, we should note that options (i) and (ii) also have the disadvantage that the environmental variables must be continuous variables (i.e. they cannot be

[2] See LP (6.2) and associated discussion.

categorical variables). If there are categorical variables, then the more complicated mixed-integer LP models, suggested by Banker and Morey (1986b), can be used.[3]

Method 4

The two-stage method involves solving a DEA problem in a first-stage analysis, involving only the traditional inputs and outputs. In the second stage, the efficiency scores from the first stage are regressed upon the environmental variables. The signs of the coefficients of the environmental variables indicate the directions of the influences, and standard hypothesis tests can be used to assess the strength of the relationships. The second-stage regression can be used to "correct" the efficiency scores for environmental factors by using the estimated regression coefficients to adjust all efficiency scores to correspond to a common level of environment (e.g. the sample means).

This method accommodates both continuous and categorical variables. It also has the advantage of being easy to calculate. All that is needed is a basic DEA program and a statistics package which can conduct ordinary least squares (OLS) regression (e.g. Excel, SHAZAM, SAS, Minitab). It should be noted, however, that frequently a significant proportion of the efficiency scores are equal to one, and that the OLS regression could predict scores greater than one. It is thus recommended that the Tobit regression method be used, because it can account for truncated data.[4] Econometrics packages such as SHAZAM and LIMDEP have commands for Tobit regression.

One disadvantage of the two-stage method is that if the variables used in the first stage are highly correlated with the second-stage variables then the results are likely to be biased.

Also note that some researchers use the above two-stage method to determine the directions of influence of the environmental variables, and then use this information to formulate a single-stage model such as that specified in Method 3(i) above.

Concluding Points on Environmental Variables

We have considered a number of possible approaches to the consideration of environmental variables. We recommend the two-stage approach in most cases. It has the advantages that:

- it can accommodate more than one variable;
- it can accommodate both continuous and categorical variables;

[3] See also Kamakura (1988) and Rousseau and Semple (1993) for further comments on mixed integer LP models.
[4] For example, see McCarty and Yaisawarng (1993).

- it does not make prior assumptions regarding the direction of the influence of the environmental variable;
- one can conduct hypothesis tests to see if the variables have a significant influence upon efficiencies;
- it is easy to calculate; and
- the method is simple and therefore transparent.

The two-stage approach can also be used to assess the influence of various management factors upon efficiency. For example, the effects upon efficiency of the age, experience, education and training of the manager(s) can be estimated by including these factors in the second-stage regression.

7.5 Input Congestion

In our discussion in Chapter 2, it was stated that a single-input production function may turn downwards and have negative slope for some values of the input. This was explained as being due to congestion in the use of the input, to the extent that it begins to have a negative marginal product. In the case of more than one input, input congestion may cause isoquants to "bend backwards" and obtain a positive slope at some point. Some DEA studies have allowed the DEA input isoquant to have segments with a positive slope to account for this possibility. It is usually argued that the excess input use is due to constraints that are not under the control of the firm, such as: labour unions preventing a reduction in staff, or government controls setting the levels of various inputs.

The standard DEA models discussed in the previous chapter implicitly assume strong disposability in inputs (and outputs). That is, it is assumed that a firm can always costlessly dispose of unwanted inputs (and outputs). A DEA model that accounts for input congestion relaxes the strong disposability in inputs assumption. In this section, we replace the assumption of strong disposability in inputs with the assumption of weak disposability in inputs.

Following Färe, Grosskopf and Lovell (1985, 1994), input congestion is accounted for in the input-orientated VRS DEA problem, defined in LP 6.12, by changing the inequalities in the input restrictions to equalities and by introducing a δ-parameter in the input restrictions. Thus, LP 6.12 becomes:

$$\begin{aligned} \min_{\theta,\lambda,\delta}\ & \theta, \\ \text{st}\quad & -\mathbf{q}_i + \mathbf{Q}\boldsymbol{\lambda} \geq \mathbf{0}, \\ & \delta\theta\mathbf{x}_i - \mathbf{X}\boldsymbol{\lambda} = \mathbf{0}, \\ & \mathbf{I1}'\boldsymbol{\lambda}=1, \\ & \boldsymbol{\lambda} \geq \mathbf{0}, \\ & 0<\delta\leq 1. \end{aligned} \tag{7.9}$$

The technical efficiency scores obtained from this weak disposability VRS DEA are greater than or equal to the strong-disposability VRS DEA scores obtained from LP 6.12. This is because the effects of congestion inefficiency have been removed from the technical efficiency measure.

Recall that, in Chapter 6, we decompose CRS technical efficiency into "pure" (VRS) technical efficiency and scale efficiency by solving separate VRS and CRS DEA models. In a similar manner, we can solve a strong-disposability and a weak disposability DEA and identify input-congestion efficiency (ICE) from the differences in the TE scores from the two models. This is illustrated in Figure 7.2 where we have a frontier constructed assuming strong disposability (SS_S) and one assuming weak disposability (SS_W). The latter curve has "bent back" to pass through the point A. The congestion inefficiency for the firm producing at P is equal to P_WP_S. In ratio terms, an input-congestion efficiency is defined by

$$\text{ICE} = 0P_S/0P_W,$$

The TE measure under strong disposability (TE_S) is equal to the product of the TE measure under weak disposability (TE_W) and the input-congestion efficiency (ICE). That is,

$$0P_S/0P = (0P_S/0P_W)\times(0P_W/0P)$$

or

$$\text{TE}_S = \text{ICE}\times\text{TE}_W.$$

Thus, by solving the strong- and weak-disposability DEA models, we can identify ICE by the ratio of TE_S to TE_W.

In fact, as shown in Färe, Grosskopf and Logan (1985), the technical efficiency scores calculated from a CRS DEA can be decomposed into congestion inefficiency, scale inefficiency and "pure" technical efficiency by solving three DEA models: a CRS assuming strong disposability; a VRS assuming strong disposability; and a VRS assuming weak disposability.

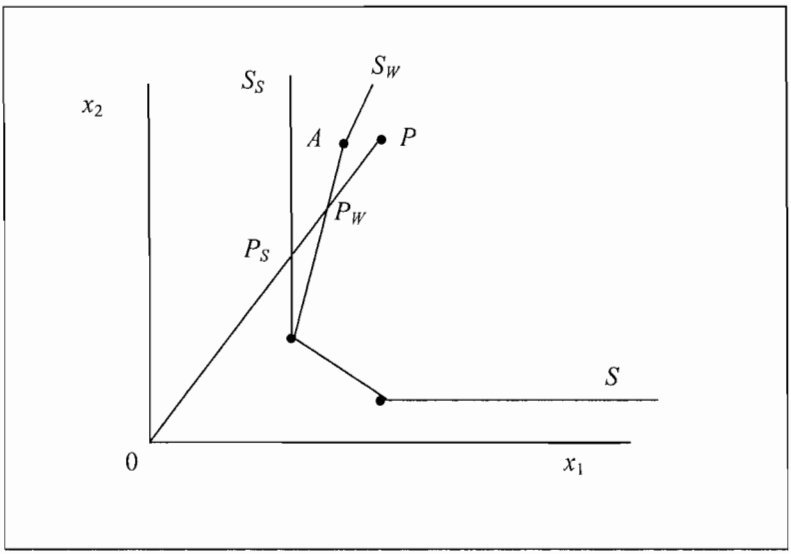

Figure 7.2 Efficiency Measurement and Input Disposability (Congestion)

Weak disposability in outputs can also be considered. This permits the existence of a positively sloped portion of the production possibility curve, implying a negative shadow price for a particular output. This allows one to explicitly include unwanted outputs such as pollutants in a DEA model. For example, see Färe, *et al.* (1989) and Färe, Grosskopf and Lovell (1985, 1994).

Note that weak disposability in both inputs and outputs can be imposed together. Furthermore, one could alternatively impose weak disposability upon a subset of the inputs and/or outputs. For example, a researcher may have a strong reason to believe that some rail companies have been required to invest in excess capital because of political directives. In that instance, the researcher may decide to impose weak disposability in the capital input and allow strong disposability in all other inputs.

We conclude this section with a word of caution. In our opinion, unless there are strong reasons for suspecting congestion, one should not go looking for it because it will often be found whether or not it actually exists. This is because the method can often identify "congestion inefficiencies" that are simply due to having insufficient data points at the extremities of the isoquants. This is also an issue that can affect the measurement of scale inefficiency.

7.6 Treatment of Slacks

As noted in Chapter 6, the LPs discussed in that chapter may not always identify all (efficiency) slacks. Some authors (e.g., Ali and Seiford, 1993) have suggested the use of a second-stage linear programming problem to ensure the identification of an efficient frontier point by maximising the sum of slacks required to move from the first-stage projected point (such as A' in Figure 6.1) to a Koopmans-efficient frontier point (such as point C in Figure 6.5). This second-stage linear programming problem is defined by:

$$\begin{aligned} &\min_{\lambda,\mathbf{OS},\mathbf{IS}} -(\mathbf{M1}'\mathbf{OS} + \mathbf{N1}'\mathbf{IS}), \\ &\text{st} \quad -\mathbf{q}_i + \mathbf{Q}\boldsymbol{\lambda} - \mathbf{OS} = \mathbf{0}, \\ &\qquad \theta\mathbf{x}_i - \mathbf{X}\boldsymbol{\lambda} - \mathbf{IS} = \mathbf{0}, \\ &\qquad \boldsymbol{\lambda} \geq \mathbf{0}, \mathbf{OS} \geq \mathbf{0}, \mathbf{IS} \geq \mathbf{0}, \end{aligned} \tag{7.10}$$

where **OS** is an $M\times1$ vector of output slacks, **IS** is a $N\times1$ vector of input slacks, and **M1** and **N1** are $M\times1$ and $N\times1$ vectors of ones, respectively. Note that in this second-stage LP, θ is not a variable, its value is taken from the first-stage results. Furthermore, note that this second-stage LP must be solved for each of the I firms involved.[5]

There are two major problems associated with this second-stage LP. First the sum of slacks is *maximised* rather than *minimised*. Hence, it identifies not the *nearest* efficient point but the *furthest* efficient point. Second, the LP is not invariant to units of measurement. The alteration of the units of measurement, say for a labour input from days to hours (while leaving other units of measurement unchanged), could result in the identification of different efficient boundary points and, hence, different slack and λ-values.[6]

However, these two issues are not a problem in the simple example that is presented in Figure 6.5 because there is only one efficient point to choose from on the vertical facet. However, if slack occurs in two or more dimensions (as is often the case) then the above-mentioned problems are relevant. Coelli (1998) suggests using a multi-stage DEA method to avoid the problems inherent in the two-stage method. This multi-stage method involves a sequence of radial DEA models and hence is more computationally demanding that the other two methods. However, the benefits of the approach are that it identifies efficient projected points which have input and output mixes as similar as possible to those of the inefficient points, and that it is also invariant to units of measurement. Hence, we recommend the use of the multi-stage method over other alternatives. For more detail on the multi-stage method, see Coelli (1998).

---

[5] This method is used in some DEA software, such as Warwick DEA and IDEAS.

[6] Charnes, *et al.* (1987) suggest a units-invariant model where the unit worth of a slack is made inversely proportional to the quantity of that input or output used by the i-th firm. This solves the immediate problem, but creates another, in that there is no obvious reason for the slacks to be weighted in this way.

In the DEAP software, the user is given three choices regarding the treatment of slacks, namely:

- One-stage DEA, in which the first-stage LP (LP 6.12) is solved and slacks are calculated residually;
- Two-stage DEA, which involves the solution of LPs 6.12 and 7.10; and
- Multi-stage DEA, which involves the solution of a sequence of radial LPs.

Having devoted some space to the issue of slacks, we conclude this discussion by observing that the importance of slacks can be overstated. Slacks may be viewed as being an artefact of the chosen frontier construction method (namely DEA) and the use of finite sample sizes. If an infinite sample size were available and/or if an alternative frontier construction method was used, which involved a smooth function surface, the slack issue would disappear. In addition to this observation, it seems quite reasonable to accept the arguments of Ferrier and Lovell (1990) that slacks may essentially be viewed as allocative inefficiency. Hence, we believe that an analysis of technical efficiency can reasonably concentrate upon the radial Farrell efficiency score provided in the first-stage DEA LP (refer to LP 6.12). However, if one insists on identifying Koopmans-efficient projected points then we strongly recommend the use of the multi-stage method in preference to the two-stage method, as outlined above.

7.7 Additional Methods

Weights Restrictions

The flexibility of the frontier that is constructed using DEA is one of the often-quoted advantages of the method, relative to parametric frontier methods. However, this aspect of the method can also create problems, especially when dealing with small data sets. In particular, one can find that the weights (i.e. shadow prices) assigned to the various input and output variables (see LP 6.2) are not realistic for some firms. They may be too large or too small (or even zero when slack regions are involved). This causes one to question the applicability of the efficiency measures obtained.

Hence, a variety of methods have been proposed which place limits upon the range of the shadow prices that may be obtained by including additional restrictions into the DEA LP. A review of commonly used methods is provided in Allen et al (1997). An early contribution was the model of Dyson and Thanassoulis (1988), in which restrictions are included into the dual LP of the form:

$$\mu \geq \mu_L, \quad \mu \leq \mu_U, \quad \nu \geq \nu_L, \quad \nu \leq \nu_U,$$

where μ_L, μ_U, ν_L, and ν_U are vectors of lower-bound and upper bound values associated with the dual weights (e.g. in LP 6.2). A number of difficulties are associated with this method, including selecting weights that correctly reflect the analyst's value judgements; ensuring that the restrictions are internally consistent; and avoiding infeasibility in the LP. Furthermore, it should be noted that the equivalence of TE scores in input- and output-orientated CRS DEA models is no longer guaranteed when these types of weights restrictions are introduced.

Various alternatives to the above types of direct weight restrictions have been tried, such as setting restrictions on the ratios of the shadow prices. That is, upon the range of values that the marginal rates of transformation (between outputs) and the marginal rates of substitution (between inputs) can take. Another alternative, proposed by Wong and Beasley (1990), involves imposing restrictions upon what they call the "virtual weights", which could be alternatively called the shadow cost shares and shadow revenue shares. For example, if one wished to impose a lower bound restriction on the shadow revenue share of the m-th output variable, one could impose a restriction of the form:

$$\mu_m q_m / \boldsymbol{\mu}'\mathbf{q} \leq SRS_{Lm},$$

where μ_m is the m-th element of $\boldsymbol{\mu}$, q_m the m-th element of $\mathbf{q}$ and SRS_{Lm} is the lower bound specified for the shadow revenue share of the m-th output. Similar restrictions (both upper and lower bound) can be imposed on each output and each input variable if required.[7] For further discussion of the relative merits of these alternative weights restrictions methods see Allen et al (1997).

Super Efficiency

The term "super efficiency" relates to an amended DEA model in which firms can obtain efficiency scores greater than one because each firm is not permitted to use itself as a peer. Consider the case of the input-orientated CRS DEA in LP 6.3. To calculate a super efficiency score for the i-th firm, the data for the i-th firm is removed from the $N \times I$ $\mathbf{X}$ matrix and from the $M \times I$ $\mathbf{Y}$ matrix so they become $N \times (I\text{-}1)$ and $M \times (I\text{-}1)$ matrices, respectively. Thus, when the LP is run the i-th firm cannot form part of its reference frontier and hence, if it was a fully-efficient frontier firm in the original standard DEA model, it may now have an efficiency score greater than one. This LP is run for each of the I firms in the sample, and, in each of these LPs, the reference set involves I-1 firms.

An illustration of this technique is provided in Figure 7.2, where five firms (A, B, C, D, E) use two inputs to produce a particular output. When the standard DEA model in LP 6.2 is applied to these data, the firms B, C and D form the frontier and, hence, each of these firms have an efficiency score of one. However, if we apply the

[7] This concept of shadow revenue and cost shares is also relevant to the discussion of TFP growth measures derived from DEA models, which is discussed in Chapter 11.

super-efficiency DEA methods it is possible for these frontier firms to obtain super efficiency scores that are greater than one. For example, consider the case of firm *C*. When measuring its super-efficiency score it will no longer form part of the frontier and, hence, the new frontier involves only two firms (*B* and *D*) and, therefore, its projected point will be *C'*. The super-efficiency score for firm *C* will be $0C'/0C$, which is approximately 1.2. This indicates that this firm could increase input usage by 20% and still be within the technology defined by the other firms in the sample. Note that the original efficiency scores of non-frontier firms (such as *A* and *E*) do not change when the super-efficiency method is used because they did not form part of the original DEA frontier.

This method was originally proposed by Andersen and Petersen (1993), who used the method to provide a ranking system that would help them discriminate between frontier firms. That is, a firm with a super-efficiency score of 1.2 is better than one with a score of 1.05 because the former is further ahead of its peers, etc. The super efficiency method has subsequently been used in a number of alternative ways. For example, in sensitivity testing, identification of outliers, and as a method of circumventing the bounded-range problem in a second stage regression method so that standard ordinary least squares regression methods can be used instead of Tobit regression. One drawback of the method is that some of the LPs may be infeasible. See Lovell and Rouse (2003) for a discussion and suggested solution of the infeasibility problem.

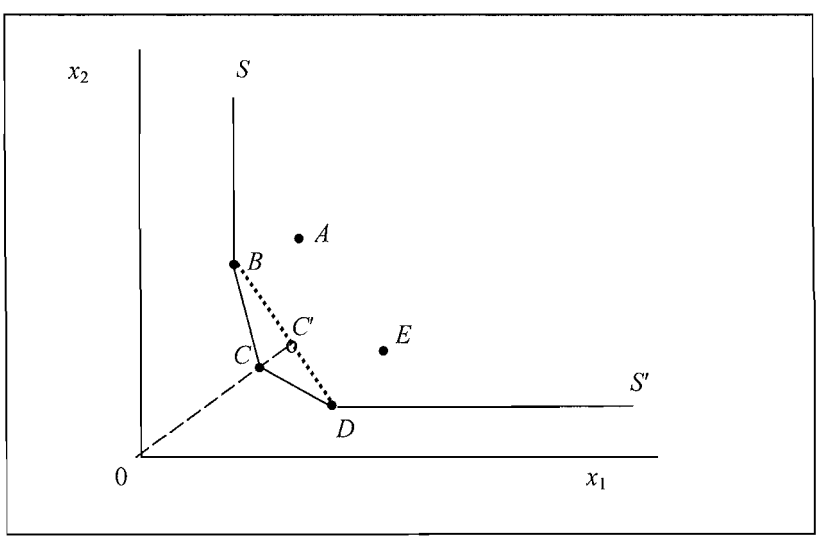

Figure 7.3 Super Efficiency

Bootstrap Methods

Here we discuss the bootstrap methods, described by Simar and Wilson (2000) and others, which attempt to provide a statistical foundation for DEA methods. The bootstrap is a re-sampling technique that has been applied in various statistical settings in recent decades, as a means of approximating the properties of the sampling distribution of an estimator (when this is difficult to obtain by using alternative means) and, hence, allowing one to conduct hypothesis tests and construct confidence intervals. See Efron and Tibshirani (1993) for information on the bootstrap.

In its simplest form, the bootstrap involves randomly selecting thousands of "pseudo samples" (using simple random sampling with replacement) from the observed set of sample data. One then obtains "pseudo estimates" from each of these samples. These thousands of pseudo estimates form an empirical distribution for the estimator of interest. This distribution is used as an approximation of the true underlying sampling distribution of the estimator.

Bootstrap methods have become easier to implement with the advent of cheap computing power, but they are still a fairly involved exercise for the average applied researcher. In the case of DEA, bootstrap methods are further complicated by the one-sided nature of the inefficiency distribution, which produces bias and inconsistency problems in certain "naive" implementations of bootstrap methods. Simar and Wilson (2000) provide a good discussion of these problems and outline some potential solutions. One solution they propose involves estimating a bias-corrected, non-parametric kernel estimate of the density of the inefficiency scores and then drawing the pseudo samples from this density. This addresses the problem with the naive bootstrap methods but does add an additional layer of complexity to the process.

Simar and Wilson (2000) also provide Monte Carlo evidence that suggests that their method works well in a simple one-input, one output case. However, they note that their methods "lack a rigorous proof of consistency" (p. 74) and they also observe that a "remaining challenge is to find a way for allowing for stochastic noise in the data in these models" (p. 75).

This latter point is a very important one, which is not well understood by some analysts. These DEA bootstrapping methods are designed to deal with *sampling variability*. That is, they provide an indication of the degree to which the efficiency estimates are likely to vary when a different sample is randomly selected from the population. However, these methods do not attempt to account for random noise, such as that which may result from measurement error or specification error. The DEA method assumes that data noise does not exist. Given this assumption (that noise does not exist), it is clear that the "height" of the DEA frontier is biased downwards in finite samples and hence that the efficiency scores are biased

upwards. However, if we assume that random noise does actually exist, then the direction of this bias could arguably be in the other direction.

Another point worth noting is that it does not make much sense for one to apply bootstrapping methods to a DEA analysis based upon census data. For example, when one has data on all cement factories in a particular country or all hospitals in a particular region. In the census case, where one has "noise-free" data on all firms in the population, the DEA frontier obtained must be the true frontier. That is, in this case the frontier has been *measured* and not *estimated*. Hence there is no need to consider sampling variability.

Overall, we have a number of reservations regarding the bootstrap because it is difficult to implement and it is not always clear that the underlying assumptions are correct. However, with these qualifications made, the bootstrap can still be useful as a way of illustrating the sensitivity of DEA efficiency estimates to variations in sample composition. In particular, the widths of bootstrap confidence intervals for the efficiency of firms located on the fringes of the data set will tend to be quite wide, indicating the degree to which these estimates are generally based upon rather thin data and, hence, should be interpreted cautiously. Furthermore, when one has a small sample and a large number of dimensions the confidence intervals tend to be wide, in general, emphasising the degree to which the method is trying to extract too much information from too few data points.

7.8 Empirical Application: Australian Universities

In 1996, senior management at the University of New England (UNE) formed a committee to look at the performance of UNE relative to other universities in Australia. Of particular concern to management was some evidence that suggested that UNE's administration sector appeared to be larger than some comparable universities. This committee commissioned a DEA study (see Coelli, 1996d) that looked at the relative performance of Australia's 36 universities. The study involved the construction of three separate models: one for the administration sectors; one for the academic sectors; and one for universities as a whole. To conserve space in this section we restrict our discussion of this study to the model for the administration sectors of Australian universities.

The data used in this analysis are for 36 Australian universities for the 1994 calendar year. An input-orientated VRS DEA model is used in the analysis. CRS and NIRS models are also run to investigate scale issues. An input orientation is chosen because we believe that, in 1994, universities had greater control over input quantities relative to output quantities (in particular, we note that the vast majority of student load was fixed by government quotas). However, one could also argue the converse, so we also ran our models assuming an output-orientation and observed that orientation had little influence on the efficiency scores obtained in the study.

Because of the limited number of observations available, Coelli (1996d) restricted the total number of input and output variables in the analysis to ensure some degree of discretionary power remained. Hence, the model involved only two inputs and two outputs. The two inputs used were *expenditure on administrative staff* and *other administrative costs*. Although physical quantity measures were preferred, these were not available for the labour input and would be too difficult to measure in the case of the *other inputs* variable. It was hypothesised that the use of the value measures were unlikely to introduce considerable bias in measures because the wage levels and the prices of other administrative inputs did not vary significantly across Australian universities in 1994.

The specification of the outputs of a university administration was a challenging task. The principal role of a university administration is to keep records on staff and students. With regard to university staff, the administration is principally involved in activities such as recruitment, payroll, study leave, and keeping financial records associated with departmental budgets and research grant budgets. With regard to the students, the administration is involved in promotion, enrolment, record keeping for fees and charges, examination records, etc. It was decided that the outputs of the university administrations are proxied by two measures: the total number of students (measured in equivalent full time student units (EFTSU)) and the total number of staff (in full-time equivalent units).

The above two-input, two-output model of university administration is an obvious simplification of reality. The variable specifications can be criticised from many angles. In an attempt to head off some of the potential criticisms, Coelli (1996d) ran the model using some different variable definitions so as to assess the sensitivity of the results to different specifications. Three alternative models were considered.

1. Because the use of EFTSU measures may underestimate the output of an administration at a university where there is a large number of part-time students, this issue was dealt with by using total enrolments instead of EFTSU. This measure, however, is likely to overstate the output of those universities (like UNE) with a large part-time student population because part-time students do fewer units per year and, hence, involve less record keeping.
2. Because the *total staff* output variable does not properly reflect the extra administrative load that results when the academic staff at a university apply for and attract a lot of research money, the *total staff* output variable was replaced with a *total research grants* variable. This variable would be appropriate if the majority of service provided by the administration staff to the staff of the university involved assisting academic staff apply for grants and the administration of grant monies obtained. This is unlikely to be true, but this analysis was conducted to see how sensitive the results were to this problem.

3. An additional model was estimated where the above two changes were simultaneously imposed.

Table 7.3 DEA Results for the Australian Universities Study

University	VRS TE	CRS TE	scale eff.	
Australian Catholic University	0.757	0.806	0.938	irs
Australian National University	1.000	1.000	1.000	-
Central Queensland University	0.499	0.557	0.896	irs
Charles Sturt University	1.000	1.000	1.000	-
Curtin University of Technology	0.700	0.702	0.997	drs
Deakin University	0.786	0.800	0.982	drs
Edith Cowan University	0.784	0.861	0.911	drs
Flinders University of South Australia	1.000	1.000	1.000	-
Griffith University	0.720	0.738	0.975	drs
James Cook University	0.725	0.757	0.958	irs
La Trobe University	0.930	0.947	0.982	drs
Macquarie University	0.770	0.778	0.990	irs
Monash University	0.728	1.000	0.728	drs
Murdoch University	0.779	0.824	0.946	irs
Northern Territory University	0.662	0.980	0.676	irs
Queensland University of Technology	0.978	1.000	0.978	drs
Royal Melbourne Institute of Tech.	0.739	0.786	0.939	drs
Southern Cross University	0.950	1.000	0.950	irs
Swinburne University of Technology	0.876	0.925	0.947	irs
University of Adelaide	0.653	0.665	0.982	drs
University of Ballarat	0.867	1.000	0.867	irs
University of Canberra	1.000	1.000	1.000	-
University of Melbourne	0.838	1.000	0.838	drs
University of New England	**0.707**	**0.713**	**0.991**	**irs**
University of New South Wales	0.745	0.930	0.801	drs
University of Newcastle	0.404	0.404	1.000	-
University of Queensland	1.000	1.000	1.000	-
University of South Australia	1.000	1.000	1.000	-
University of Southern Queensland	0.798	0.826	0.967	irs
University of Sydney	0.765	1.000	0.765	drs
University of Tasmania	1.000	1.000	1.000	-
University of Technology, Sydney	1.000	1.000	1.000	-
University of Western Australia	0.865	0.870	0.995	irs
University of Western Sydney	0.622	0.625	0.995	drs
University of Wollongong	0.882	0.892	0.989	irs
Victoria University of Technology	0.904	0.918	0.985	irs
Mean	**0.818**	**0.870**	**0.944**	

The DEA results of the (base-run) university administration model are listed in Table 7.3. We observe that UNE achieves a (VRS) technical efficiency score of 0.713 which ranks it 31st in 36 universities. This result appears to confirm the suspicions of the UNE management. The average technical efficiency score in the sample is 0.818 with the lowest technical efficiency score being 0.404 for Newcastle. The UNE scale efficiency score is 0.991, which indicates that UNE is

operating quite close to optimal scale. From the computer printout (which is not listed) it was observed that UNE's peers were Flinders, Tasmania, Ballarat and the Australian National University (ANU), in order of importance. The relative weights of these peers were: 0.49, 0.41, 0.09 and 0.01, respectively. Hence Flinders and Tasmania provide 90% of the peer weighting. This information is reassuring because Flinders and Tasmania have quite similar structures to UNE.

The three alternative DEA model specifications discussed above were also run to test the robustness of the results. The results did improve slightly for UNE (because the changes were particularly structured to do so) but it was still apparent that even under these favourable assumptions the UNE administration was still observed to have significant room for improvement.

UNE is one of the principal providers of distance education in Australia. It enrols approximately 5,000 on-campus students and 14,000 off-campus (distance education) students each year. Some observers argue that distance education students are more costly in terms of administrative requirements. In an attempt to address this issue the two-stage method, described in Section 7.3, was applied. This involved the estimation of a second-stage Tobit regression in which the VRS technical efficiency scores from the base model were regressed upon the percentage of external enrolments in each university and the average unit load of students (measured by the ratio of EFTSU to total enrolments). The results suggested that neither of these factors had a significant influence upon the efficiency scores, with the Tobit regression equation explaining less than 3% of the total variation in the technical efficiency scores.

7.9 Conclusions

In this chapter we discuss a limited number of extensions to the basic DEA models outlined in Chapter 6. A discussion of all the extensions that have been proposed in the literature is not possible here. Some additional model extensions that could be pursued include: the stochastic DEA models proposed by Land, Lovell and Sten (1993) and Olsen and Petersen (1995); the additive model proposed by Charnes, Cooper, Golany, Seiford and Stutz (1985); the *Flexible Disposable Hull* (FDH) approach of Deprins, Simar and Tulkens (1984), which relaxes convexity assumptions; and the development of panel data methods, such as the window analysis method proposed by Charnes, Clark, Cooper and Golany (1985) and the Malmquist index approach of Färe, Grosskopf, Norris and Zhang (1994). The Malmquist method is discussed in some detail in Chapter 11.

For further reading on DEA methods, there are a number of useful books available. These include Färe, Grosskopf and Lovell (1985, 1994), Ganley and Cubbin (1992), Fried, Lovell and Schmidt (1993), Charnes et al (1995), Färe and Grosskopf (1996), Cooper, Seiford and Tone (2000) and Thanassoulis (2001).

Before finishing this discussion of DEA, it is important to list a few of the limitations and possible problems that a researcher may encounter in conducting a DEA study. These include:

- Measurement error and other noise may influence the shape and position of the frontier.
- Outliers may influence the results.
- The exclusion of an important input or output can result in biased results.
- The efficiency scores obtained are only relative to the best firms in the sample. The inclusion of extra firms (e.g., from other countries) may reduce efficiency scores.
- Be careful when comparing the mean efficiency scores from two studies. They only reflect the dispersion of efficiencies within each sample – they say nothing about the efficiency of one sample relative to the other.
- The addition of an extra firm in a DEA analysis cannot result in an increase in the TE scores of the existing firms.[8]
- The addition of an extra input or output in a DEA model cannot result in a reduction in the TE scores.
- When one has few observations and many inputs and/or outputs many of the firms will appear on the DEA frontier.[9]
- Treating inputs and/or outputs as homogenous commodities when they are heterogenous may bias results.
- Not accounting for environmental differences may give misleading indications of relative managerial competence.
- Standard DEA does not account for multi-period optimisation nor risk in management decision making.[10]

---

[8] See Zhang and Bartels (1998) for an illustration of the effect of sample size upon mean TE scores in DEA.

[9] This point and the previous two points are related. One implication is that if an investigator wished to make an industry look good, he/she could reduce the sample size and increase the number of inputs and outputs in order to increase the TE scores.

[10] See Färe and Grosskopf (1996) for discussion of some dynamic DEA models.

It should also be stressed that all of these issues are also applicable (in varying degrees) to the stochastic frontier method discussed in the following chapters. The relative merits of the two approaches are discussed further in Chapter 12.

8. ECONOMETRIC ESTIMATION OF PRODUCTION TECHNOLOGIES

8.1 Introduction

In the last two chapters we have seen how DEA can be used to estimate the position and shape of the production frontier introduced in Chapter 2. The DEA method is computationally simple and has the advantage that it can be implemented without knowing the algebraic form of the relationship between outputs and inputs (i.e., we can estimate the frontier without knowing whether output is a linear, quadratic, exponential or some other function of inputs). Stochastic frontier analysis (SFA) is an alternative method for frontier estimation that assumes a given functional form for the relationship between inputs and an output. When the functional form is specified then the unknown parameters of the function need to be estimated using econometric techniques. These requirements make SFA more computationally demanding than DEA. However, as we see in Chapter 9, SFA has some advantages that make the extra computational burden worthwhile.

As a pre-cursor to a detailed discussion of stochastic frontiers, this chapter provides an overview of econometric methods for estimating economic relationships. Because this book is mainly concerned with production relationships, our discussion is couched in terms of the production and cost functions introduced in Chapter 2. In addition, to maintain consistency of notation between this chapter and the discussion of stochastic frontiers in Chapter 9, we focus mainly on the case where we have access to cross-sectional data (i.e., observations on several firms at a single point in time). However, the econometric concepts and methods we discuss carry over, often with little or no modification, to cases where we have time-series

data (i.e., observations on a single firm at several points in time) or panel data (i.e., observations on several firms at several points in time).

We begin, in Section 8.2, by listing some of the more common functional forms found in the applied economics literature, including the linear, Cobb-Douglas, normalised quadratic and translog functional forms. In Section 8.3, we describe how the parameters of different models can be estimated using the methods of ordinary least squares, maximum likelihood and nonlinear least squares. In Section 8.4, we show how these methods can be adapted to ensure the parameters of our models satisfy equality constraints implied by economic theory (eg., homogeneity and symmetry constraints) and, in Section 8.5, we show how to test the validity of these constraints using F- and likelihood-ratio tests. At this point, readers will have developed enough econometrics background to understand the basic stochastic frontier models to be discussed in Chapter 9.

In the remainder of the chapter, we discuss material intended for readers who will be using the more advanced stochastic frontier methods discussed in Chapter 10. In Section 8.6, we discuss the advantages of estimating several seemingly unrelated economic relationships as part of a system. In Section 8.7, we present examples of how economic theory can give rise to inequality constraints involving the parameters (eg., concavity constraints) and show how *some* types of inequality constraints can be handled using nonlinear least squares. In Section 8.8, we show how to estimate models and impose inequality constraints using the Bayesian approach to inference, an approach that usually requires the evaluation of analytically intractable integrals. In Section 8.9, we show how these integrals can be evaluated using simulation methods.

Throughout this chapter, we illustrate different estimation and hypothesis testing procedures using the Philippine rice data that are discussed in Appendix 2.

8.2 Production, Cost and Profit Functions

The production, cost and profit functions discussed in Chapter 2 express a single dependent variable as a function of one or more explanatory variables. For example, a production function expresses one output as a function of inputs, while a variable cost function expresses cost as a function of input prices and outputs. Mathematically, all these different functions can be written in the form:

$$y = f(x_1, x_2, ..., x_N) \tag{8.1}$$

where y is the dependent variable; the x_n (n = 1, ..., N) are explanatory variables; and $f(.)$ is a mathematical function about which economic theory usually has little or nothing to say. Thus, the first step in estimating the relationship between the dependent and explanatory variables is to specify the algebraic form of $f(.)$.

8.2.1 Common Functional Forms

Different algebraic forms of $f(.)$ give rise to different models. Some common functional forms are listed in Table 8.1, where γ and the β_n and β_{mn} are unknown parameters to be estimated[1]. When choosing between these different forms, we usually give preference to those that are:

1. *Flexible.* A functional form is said to be *first-order flexible* if it has enough parameters to provide a first-order differential approximation to an arbitrary function at a single point.[2] A *second-order flexible* form has enough parameters to provide a second-order approximation. The linear and Cobb-Douglas forms are first-order flexible, while the remaining functional forms listed in Table 8.1 are second-order flexible. All other things being equal, we usually prefer functional forms that are second-order flexible. However, increased flexibility comes at a cost – there are more parameters to estimate, and this may give rise to econometric difficulties (eg., multicollinearity).

Table 8.1 Some Common Functional Forms

Linear	$y = \beta_0 + \sum_{n=1}^{N} \beta_n x_n$
Cobb-Douglas	$y = \beta_0 \prod_{n=1}^{N} x_n^{\beta_n}$
Quadratic	$y = \beta_0 + \sum_{n=1}^{N} \beta_n x_n + \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \beta_{nm} x_n x_m$
Normalised quadratic	$y = \beta_0 + \sum_{n=1}^{N-1} \beta_n \left(\frac{x_n}{x_N} \right) + \frac{1}{2} \sum_{n=1}^{N-1} \sum_{m=1}^{N-1} \beta_{nm} \left(\frac{x_n}{x_N} \right) \left(\frac{x_m}{x_N} \right)$
Translog	$y = \exp\left(\beta_0 + \sum_{n=1}^{N} \beta_n \ln x_n + \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \beta_{nm} \ln x_n \ln x_m \right)$
Generalised Leontief	$y = \sum_{n=1}^{N} \sum_{m=1}^{N} \beta_{nm} \left(x_n x_m \right)^{1/2}$
Constant Elasticity of Substitution (CES)	$y = \beta_0 \left(\sum_{n=1}^{N} \beta_n x_n^{\gamma} \right)^{1/\gamma}$

[1] The β_{mn} parameters satisfy the identifying condition $\beta_{mn} = \beta_{nm}$ for all n and m. This condition is sometimes known as a *symmetry* condition. Note that some functional forms are special cases of others. For example, the Cobb-Douglas can be obtained from the translog by setting all $\beta_{mn} = 0$.

[2] The phrase "n-th order differential approximation to an arbitrary function at a single point" means it is possible to choose values of the parameters so that the value of the approximating function and all its derivatives up to order n are equal to those of the arbitrary function at that point.

2. *Linear in the parameters.* Many of the functions listed in Table 8.1 are linear in the parameters, making them amenable to estimation using the linear regression techniques that are discussed later in this chapter. At first glance, the Cobb-Douglas and translog functions appear not to satisfy this property. However, taking the logarithms of both sides of these functions yields

Cobb-Douglas: $$\ln y = A_0 + \sum_{n=1}^{N} \beta_n \ln x_n \text{ where } A_0 = \ln \beta_0 \tag{8.2}$$

Translog: $$\ln y = \beta_0 + \sum_{n=1}^{N} \beta_n \ln x_n + \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \beta_{nm} \ln x_n \ln x_m \tag{8.3}$$

which are both linear in the parameters. Thus, the parameters of Cobb-Douglas and translog functions can also be estimated in a linear regression framework.

3. *Regular.* Some of the functions listed in Table 8.1 automatically satisfy the economic regularity properties discussed in Chapter 2. For example, the normalised quadratic function is automatically homogeneous of degree zero (this makes it particularly suitable for the estimation of demand and supply relationships). In other cases, it is usually possible to devise and impose simple restrictions on the parameters that are sufficient for certain properties to be satisfied. For example, a translog function is homogeneous of degree r if

$$\sum_{n=1}^{N} \beta_n = r \quad \text{and} \quad \sum_{n=1}^{N} \beta_{nm} = \sum_{m=1}^{N} \beta_{nm} = 0, \tag{8.4}$$

and a generalised Leontief function is concave if

$$\beta_{nm} \geq 0 \text{ for all } n \neq m. \tag{8.5}$$

4. *Parsimonious.* The principle of parsimony says we should choose the simplest functional form that "gets the job done adequately". Sometimes we can assess the adequacy of a functional form prior to estimation. For example, the Cobb-Douglas function is inadequate in situations where elasticities may vary across data points,[3] and both the Cobb-Douglas and translog functions are problematic when the data contain zeros because this makes it impossible to construct the logarithms of the variables.[4] However, model adequacy is often determined after estimation by conducting a residual analysis (i.e., assessing whether residuals exhibit any systematic patterns that are indicative of a poorly chosen function), hypothesis testing, calculating measures of goodness-of-fit and assessing predictive performance.

More information on functional forms is available in Fuss, McFadden and Mundlak (1978) and Chambers (1988).

---

[3] The Cobb-Douglas elasticities are constant. Other restrictive properties are that the returns to scale is a constant and the elasticity of substitution is unity – see Chapter 2.

[4] A method for handling zero input observations has been proposed by Battese (1997).

8.2.2 Accounting for Technological Change

Technological advances often cause economic relationships (especially production functions) to change over time. If we have observations over time, we usually account for technological change by including a time trend in our model. For example, the following models all account for technological change:

Linear: $$y = \beta_0 + \theta t + \sum_{n=1}^{N} \beta_n x_n \tag{8.6}$$

Cobb-Douglas: $$\ln y = A_0 + \theta t + \sum_{n=1}^{N} \beta_n \ln x_n \tag{8.7}$$

Translog: $$\ln y = \beta_0 + \theta_1 t + \theta_2 t^2 + \sum_{n=1}^{N} \beta_n \ln x_n + \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \beta_{nm} \ln x_n \ln x_m \tag{8.8}$$

where t is a time trend; and θ, θ_1 and θ_2 are unknown parameters to be estimated.

When economists include time trends in their models they are making implicit assumptions about the nature of technological change. To see this, take the specifications 8.6 to 8.8 and consider the percentage change in y in each period due to technological change. This is given by the derivative of $\ln y$ with respect to t:

Linear: $$\frac{\partial \ln y}{\partial t} = \frac{\theta}{y} \tag{8.9}$$

Cobb-Douglas: $$\frac{\partial \ln y}{\partial t} = \theta \tag{8.10}$$

Translog: $$\frac{\partial \ln y}{\partial t} = \theta_1 + 2\theta_2 t \tag{8.11}$$

Thus, the linear specification implicitly assumes the technological change effect is inversely related to y, the Cobb-Douglas specification implicitly assumes it is constant, while the translog model allows the technological change effect to increase or decrease with time (depending on whether θ_2 is positive or negative). In practice, the way we introduce a time trend into our model should reflect our industry-specific knowledge of technological developments and how they impact on important characteristics of economic behaviour. For example, in a production context, we might ask ourselves whether technological change merely increases average output, or whether it also changes marginal rates of technical substitution.[5] In the latter case, we should introduce t into the model in such a way that it allows some of the slope coefficients to change over time (an example is presented in Chapter 11). For more details see Chambers (1988).

[5] If the marginal rate of technical substitution is independent of time then technical change is said to be *Hicks neutral*.

8.3 Single Equation Estimation

A possible source of error in econometric estimation is the inadvertent omission of relevant variables from the right-hand side of the economic relationship 8.1. Other sources of error include errors of measurement and the approximation errors introduced when the function $f(.)$ in equation 8.1 is approximated by functional forms such as those discussed above. We can account for the combined effects of these types of errors (known as *statistical noise*) by including random variables in our models. For the time being, we focus on models that are linear in the parameters and write

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + v_i \qquad i = 1, \ldots, I, \tag{8.12}$$

where y_i denotes the i-th observation on the dependent variable; $\mathbf{x}_i$ is a $K \times 1$ vector containing the explanatory variables; $\boldsymbol{\beta}$ is an associated $K \times 1$ vector of unknown parameters; v_i is a random error term representing statistical noise; and I denotes the number of observations in the data set. Depending on the functional form chosen and the type of economic relationship being estimated, the dependent and explanatory variables in the model 8.12 may be different functions (eg., logarithms, square roots, ratios) of input and output prices and quantities. For example, if we choose a translog model to explain variations in an output q_i as a function of inputs x_{1i}, x_{2i} and x_{3i} then

$$y_i = \ln q_i\,, \tag{8.13}$$

$$\mathbf{x}_i = \begin{bmatrix} 1 \\ \ln x_{1i} \\ \ln x_{2i} \\ \ln x_{3i} \\ 0.5(\ln x_{1i})^2 \\ \ln x_{1i} \ln x_{2i} \\ \ln x_{1i} \ln x_{3i} \\ 0.5(\ln x_{2i})^2 \\ \ln x_{2i} \ln x_{3i} \\ 0.5(\ln x_{3i})^2 \end{bmatrix}, \tag{8.14}$$

and $$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 & \beta_1 & \beta_2 & \beta_3 & \beta_{11} & \beta_{12} & \beta_{13} & \beta_{22} & \beta_{23} & \beta_{33} \end{pmatrix}'. \tag{8.15}$$

Having written our model in the compact form 8.12, our next task is to estimate the unknown parameter vector $\boldsymbol{\beta}$. The two main estimation methods, ordinary least squares and maximum likelihood, are underpinned by important assumptions concerning the error terms.

8.3.1 Ordinary Least Squares (OLS) Estimation

The most common assumptions made concerning the errors are:

$$E(v_i) = 0, \qquad \text{(zero mean)} \qquad (8.16)$$

$$E(v_i^2) = \sigma^2, \qquad \text{(homoskedastic)} \qquad (8.17)$$

and $$E(v_i v_s) = 0 \text{ for all } i \neq s. \qquad \text{(uncorrelated)} \qquad (8.18)$$

The regression equation 8.12 and assumptions 8.16 to 8.18 are collectively known as the *classical linear regression model.* Technically, this model also includes the assumption that the variables in $\mathbf{x}_i$ are not random and are not exact linear functions of each other (i.e., not perfectly *collinear*).

The least squares approach to estimating $\boldsymbol{\beta}$ involves minimising the sum of squared deviations between the y_is and their means. The function that expresses this sum of squares as a function of $\boldsymbol{\beta}$ is

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{I} \left(y_i - E[y_i]\right)^2 = \sum_{i=1}^{I} \left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2. \qquad (8.19)$$

Maximising this function with respect to $\boldsymbol{\beta}$ is a straightforward exercise in calculus and involves setting the vector of first-order derivatives to zero. The solution to these first-order conditions is the *ordinary least squares (OLS)* estimator

$$\mathbf{b} = \left(\sum_{i=1}^{I} \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sum_{i=1}^{I} \mathbf{x}_i y_i. \qquad (8.20)$$

The OLS estimator is a random variable – if it were possible to collect many (random) samples of size I on the variables in the model, we could use the formula 8.20 to compute many estimates of $\boldsymbol{\beta}$. In this repeated sampling context, it can be shown that the OLS estimator has a mean of $\boldsymbol{\beta}$ (i.e., the estimator is *unbiased*). The OLS estimator is also a linear function of the y_i observations and can be shown to have a variance that is no larger than the variance of any other linear unbiased estimator (i.e., it is *efficient*). We summarise these properties by saying that the OLS estimator is the *best linear unbiased estimator (BLUE)* of $\boldsymbol{\beta}$.

More details concerning OLS estimation, including the formulas used to compute OLS standard errors, are available in econometrics textbooks such as Hill, Griffiths and Judge (2001) and Greene (2003). In practice, we do the computations using computer software packages such as EViews and SHAZAM. To illustrate, Table 8.2 presents annotated SHAZAM output from the estimation of a three-input translog production function. The model is given by equations 8.13 to 8.15 and the

data are a subset of the Philippine rice data described in Appendix 2.[6] For simplicity and ease of interpretation, we only use three inputs (area, labour and fertiliser) and all variables are rescaled to have unit means.[7] Lines in Table 8.2 beginning with the symbol "|_" are SHAZAM command lines.

Table 8.2 OLS Estimation of a Translog Production Function

```
|_ols lq lx1-lx3 lx11-lx13 lx22-lx23 lx33

 REQUIRED MEMORY IS PAR=      156 CURRENT PAR=    28000
  OLS ESTIMATION
       344 OBSERVATIONS      DEPENDENT VARIABLE= LQ
...NOTE..SAMPLE RANGE SET TO:        1,     344

 R-SQUARE =    0.8709      R-SQUARE ADJUSTED =    0.8674
VARIANCE OF THE ESTIMATE-SIGMA**2 =  0.10188
STANDARD ERROR OF THE ESTIMATE-SIGMA =  0.31919
SUM OF SQUARED ERRORS-SSE=    34.029
MEAN OF DEPENDENT VARIABLE = -0.32631
LOG OF THE LIKELIHOOD FUNCTION = -90.2070

VARIABLE    ESTIMATED   STANDARD   T-RATIO        PARTIAL STANDARDIZED ELASTICITY
  NAME     COEFFICIENT   ERROR     334 DF    P-VALUE CORR. COEFFICIENT  AT MEANS
LX1        0.57777     0.8513E-01  6.787     0.000 0.348     0.5274     0.4814
LX2        0.18002     0.8075E-01  2.229     0.026 0.121     0.1634     0.1529
LX3        0.21657     0.5043E-01  4.294     0.000 0.229     0.2398     0.2707
LX11      -0.46265     0.2475     -1.869     0.062-0.102    -0.2878     0.5050
LX12       0.68621     0.2171      3.161     0.002 0.170     0.7844    -1.3974
LX13       0.69761E-01 0.1459      0.4781    0.633 0.026     0.0941    -0.1669
LX22      -0.73097     0.3039     -2.405     0.017-0.130    -0.4195     0.7927
LX23      -0.18651     0.1389     -1.343     0.180-0.073    -0.2466     0.4509
LX33       0.27510E-01 0.9805E-01  0.2806    0.779 0.015     0.0249    -0.0466
CONSTANT   0.13938E-01 0.2471E-01  0.5642    0.573 0.031     0.0000    -0.0427
```

Approximately 87% of the variation in the logarithm of output is explained by the model.

The elasticity of output with respect to area is estimated to be 0.58 (when evaluated at the variable means).

The R^2 value reported in Table 8.2 suggests that the model explains 87% of the variation in the dependent variable. For details on measures of goodness-of-fit and their interpretation, see Hill, Griffiths and Judge (2001) and Greene (2003). The variables have been rescaled to have unit means so the first-order coefficients can be interpreted as elasticities of output with respect to inputs (when evaluated at the variable means). The estimated standard errors are generally small relative to the estimated coefficients indicating that the unknown parameters are non-zero – formal tests of hypotheses concerning the unknown parameters are discussed in Section 8.5.

8.3.2 Maximum Likelihood (ML) Estimation

The concept of maximum likelihood (ML) estimation is underpinned by the idea that a particular sample of observations is more likely to have been generated from some distributions than from others. For example, if the mean of a sample of observations is $\bar{y} = 5.3$ then, all other things being equal, the sample is more likely to have been

[6] The Philippine data comprises observations on 43 farmers over 8 years. In this section, we ignore the panel nature of the observations – we pool the data under the assumption that all firms have access to the same technology in every period, and there are zero covariances between any error terms.

[7] For example, $q_i = \text{PROD}_i / \overline{\text{PROD}}$ and $x_{1i} = \text{AREA}_i / \overline{\text{AREA}}$ where overbars denote variable means.

generated from a distribution with mean $\mu = 5$ than from a distribution with $\mu = 100$. The *maximum likelihood estimate* of an unknown parameter is defined to be the value of the parameter that maximises the probability (or likelihood) of randomly drawing a particular sample of observations.

To use the maximum likelihood principle to estimate the parameters of the classical linear regression model, we first need to make an assumption concerning the distributions of the error terms. The most common assumption is that the errors are normally distributed. Combining this assumption with assumptions 8.16 to 8.18, we write

$$v_i \sim iidN(0, \sigma^2) \tag{8.21}$$

which says the errors are independently and identically distributed normal random variables with zero means and variances σ^2. Using the relationship between y_i and v_i given by the model 8.12, together with well-known properties of normal random variables, we can then write

$$y_i \sim iidN(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2). \tag{8.22}$$

Moreover, we can write the joint density function for the vector of observations $\mathbf{y} = (y_1, y_2, ..., y_I)'$ as

$$L(\mathbf{y} \mid \boldsymbol{\beta}, \sigma) = \left(2\pi\sigma^2\right)^{-I/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{I}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2\right\}. \tag{8.23}$$

This joint probability density function (pdf) is known as the *likelihood function*. It expresses the likelihood of observing the sample observations as a function of the unknown parameters $\boldsymbol{\beta}$ and σ^2. The ML estimator of $\boldsymbol{\beta}$ is obtained by maximising this function with respect to $\boldsymbol{\beta}$. Equivalently, it can be obtained by maximising the logarithm of the likelihood function,[8]

$$\ln L = -\frac{I}{2}\ln(2\pi) - \frac{I}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{I}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2. \tag{8.24}$$

Maximising this so-called *log-likelihood function* with respect to $\boldsymbol{\beta}$ is another simple exercise in differential calculus. In fact, since the last term on the right-hand side of 8.24 is the sum of squares function 8.19, maximising the log-likelihood with respect to $\boldsymbol{\beta}$ is equivalent to minimising the sum of squares. Thus, in the special case of the classical linear regression model with normally distributed errors, the ML estimator for $\boldsymbol{\beta}$ is identical to the OLS estimator.

[8] The logarithmic transformation is a monotonic transformation, so any parameter values that maximise the likelihood function also maximise the log-likelihood function.

ML estimators are popular in empirical work because, irrespective of the type of model being estimated, if the assumptions underlying the model are valid then the ML estimator has several desirable large-sample (i.e., *asymptotic*) properties. Specifically, the ML estimator can be shown to be consistent[9] and asymptotically normally distributed (CAN) with variances that are no larger than the variances of any other CAN estimator (i.e., the ML estimator is *asymptotically efficient*).

8.3.3 Estimation of Nonlinear Models

The linear regression framework discussed above can be generalised to models that are nonlinear in the parameters. Such models can be written in the compact form

$$y_i = h(\mathbf{x}_i, \boldsymbol{\beta}) + v_i \tag{8.25}$$

where $h(.)$ is a known function; $\mathbf{x}_i$ is (still) a vector of explanatory variables; and $\boldsymbol{\beta}$ is (still) a vector of unknown parameters (but these two vectors no longer need to have the same numbers of elements). For example, in the case of a CES production function involving three inputs,

$$\mathbf{x}_i = \begin{pmatrix} x_{1i} & x_{2i} & x_{3i} \end{pmatrix}', \tag{8.26}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 & \beta_1 & \beta_2 & \beta_3 & \gamma \end{pmatrix}', \tag{8.27}$$

and $$h(\mathbf{x}_i, \boldsymbol{\beta}) = \beta_0 \left(\beta_1 x_{1i}^{\gamma} + \beta_2 x_{2i}^{\gamma} + \beta_3 x_{3i}^{\gamma} \right)^{1/\gamma}. \tag{8.28}$$

If we maintain our assumptions concerning the error terms, we can still estimate $\boldsymbol{\beta}$ using the least squares or maximum likelihood principles. We simply replace $\mathbf{x}_i'\boldsymbol{\beta}$ with $h(\mathbf{x}_i, \boldsymbol{\beta})$ in the sum of squares function 8.19 or the log-likelihood function 8.24, take first-derivatives with respect to $\boldsymbol{\beta}$ and set them to zero. In both cases, the solution to the first-order conditions is the *nonlinear least squares (NLS)* estimator. Unfortunately, the first-order conditions are usually highly nonlinear and cannot be solved analytically for convenient formulas like 8.20. Rather, we must compute the NLS estimates using an iterative optimisation procedure that involves systematically evaluating the sum of squares (or log-likelihood) function at different values of the parameters until the sum-of-squares-minimising (or log-likelihood-maximising) values are found. Details concerning these optimisation procedures are available in Greene (2003) and Judge *et al* (1985). Fortunately, the computations can be done easily using well-known econometrics software packages. To illustrate, Table 8.3 presents annotated SHAZAM output from the estimation of the three-input CES production function above. The data are the Philippine rice data used to estimate the translog production function in Section 8.3.1.

[9] An estimator is consistent, for a scalar parameter, if its values approach the true parameter value and if its variances get smaller as the sample size increases indefinitely.

Table 8.3 NLS Estimation of a CES Production Function

```
|_nl 1/ ncoef = 5
...NOTE..SAMPLE RANGE SET TO:      1,    344
|_eq q = b0*(b1*x1**g + b2*x2**g + b3*x3**g)**(1/g)
|_coef  b0 .01 b1 .4 b2 .4 b3 .1 g 1
   4 VARIABLES IN  1 EQUATIONS WITH   5 COEFFICIENTS
      344 OBSERVATIONS

REQUIRED MEMORY IS PAR=     168 CURRENT PAR=   28000

COEFFICIENT STARTING VALUES
B0        0.10000E-01  B1        0.40000      G          1.0000
B2        0.40000      B3        0.10000
      100 MAXIMUM ITERATIONS, CONVERGENCE =   0.100000E-04

INITIAL STATISTICS :

TIME =      0.016 SEC.   ITER. NO.     0   FUNCT. EVALUATIONS     1
LOG-LIKELIHOOD FUNCTION=   -566.9244
COEFFICIENTS
  0.1000000E-01  0.4000000       1.000000       0.4000000      0.1000000
GRADIENT
   289.4807       3.169468      0.3395198       3.220340       3.388841


INTERMEDIATE STATISTICS :

TIME =      0.217 SEC.   ITER. NO.    15   FUNCT. EVALUATIONS    29
LOG-LIKELIHOOD FUNCTION=   -89.57228
COEFFICIENTS
   1.103599      0.4979090      -2.498002       0.5020802      0.1225000

GRADIENT
 -0.4630688E-02  0.3881987E-03  0.2471915E-03  0.3621600E-02  0.2790431E-03


FINAL STATISTICS :

TIME =      0.237 SEC.   ITER. NO.    19   FUNCT. EVALUATIONS    33
LOG-LIKELIHOOD FUNCTION=   -89.57228
COEFFICIENTS
   1.103595      0.4978823      -2.497725       0.5020887      0.1225045

GRADIENT
  0.4598767E-04 -0.2560025E-04 -0.2710389E-05 -0.1499494E-04 -0.3633842E-06


MAXIMUM LIKELIHOOD ESTIMATE OF SIGMA-SQUARED =   0.98558E-01
SUM OF SQUARED ERRORS =    33.904
GTRANSPOSE*INVERSE(H)*G  STATISTIC  -  =  0.63545E-11

      COEFFICIENT    ST. ERROR    T-RATIO

B0         1.1036     0.18759E-01  58.829
B1        0.49788     0.90418E-01  5.5065
G         -2.4977      1.3129     -1.9024
B2        0.50209     0.90032E-01  5.5768
B3        0.12250     0.52994E-01  2.3117
|_end
```

Starting values are arbitrary (but should be plausible).

Maximised value of log-likelihood function

After 19 iterations the estimated gradients (i.e., first-order derivatives of the log-likelihood function with respect to the parameters) are all close to zero – this suggests we have reached a maximum.

Unfortunately, iterative optimisation procedures are not foolproof – sometimes they take us to a local rather than a global maximum, and sometimes they do not converge (to a maximum) at all. To confirm that the procedure has converged, it is important to check that the first derivatives of the log-likelihood function (i.e., the gradients) are close to zero. To confirm that the maximum is a global maximum, we should try and establish that the log-likelihood function is globally concave. Alternatively, we can simply re-estimate the model using different sets of starting values – if very different starting values yield the same ML estimates then it is likely that a global maximum has been reached.

Unlike the OLS estimator, the NLS estimator will not, in general, be a linear function of the y_is. This makes it difficult or impossible to establish its properties in small samples. However, the NLS estimator can be shown to be consistent and asymptotically normally distributed (whether or not the errors in the model are normally distributed). For more details on the estimation of nonlinear models see Judge *et al.* (1982, p. 633ff).

8.4 Imposing Equality Constraints

Sometimes economic theory provides us with information about the unknown parameters over and above the information contained in the data. Examples include the regularity conditions discussed in Chapter 2, many of which can be expressed in the form of equality constraints involving the parameters. Our specialist knowledge of an industry may also give rise to these types of constraints. For example, our knowledge of smallholder rice production may lead us to believe the technology exhibits constant returns to scale (CRS). In the case of the translog production function model 8.13 to 8.15, this property implies

$$\beta_1 + \beta_2 + \beta_3 = 1, \tag{8.29}$$
$$\beta_{11} + \beta_{12} + \beta_{13} = 0, \tag{8.30}$$
$$\beta_{12} + \beta_{22} + \beta_{23} = 0, \tag{8.31}$$
and
$$\beta_{13} + \beta_{23} + \beta_{33} = 0. \tag{8.32}$$

The least squares and maximum likelihood estimation methods discussed in Section 8.3 can both be adapted to enforce these types of constraints on estimates for $\boldsymbol{\beta}$. We simply minimise the sum of squares function or maximise the likelihood function subject to the constraints. Both of these constrained optimisation problems involve setting up a Lagrangean function, equating the first-order derivatives to zero and solving for $\boldsymbol{\beta}$. The method we use to solve these first-order conditions depends on the form of the model and the constraints:

a) *If both the regression model and the constraints are linear in the parameters* the first-order conditions can be solved analytically for a linear estimator known as the *restricted least squares (RLS)* estimator. If the restrictions are true the RLS estimator is the BLUE of $\boldsymbol{\beta}$. However, if the restrictions are not true the RLS estimator is biased.

   Most computer packages have RESTRICT commands that compute the RLS estimates automatically. However, if our software doesn't have this facility, we can compute them by substituting the constraints into the original model and applying OLS to the result. To illustrate the method, consider the problem of imposing the constraints 8.29 to 8.32 on the parameters of the translog model 8.13 to 8.15. This model can be written

$$\begin{aligned}\ln q_i = {} & \beta_0 + \beta_1 \ln x_{1i} + \beta_2 \ln x_{2i} + \beta_3 \ln x_{3i} \\ & + 0.5\beta_{11}\left(\ln x_{1i}\right)^2 + \beta_{12}\left(\ln x_{1i} \ln x_{2i}\right) + \beta_{13}\left(\ln x_{1i} \ln x_{3i}\right) \\ & + 0.5\beta_{22}\left(\ln x_{2i}\right)^2 + \beta_{23}\left(\ln x_{2i} \ln x_{3i}\right) \\ & + 0.5\beta_{33}\left(\ln x_{3i}\right)^2 + v_i \,. \end{aligned} \tag{8.33}$$

The constraints can also be rewritten as

$$\beta_1 = 1 - \beta_2 - \beta_3, \tag{8.34}$$
$$\beta_{11} = -\beta_{12} - \beta_{13}, \tag{8.35}$$
$$\beta_{22} = -\beta_{12} - \beta_{23}, \tag{8.36}$$

and

$$\beta_{33} = -\beta_{13} - \beta_{23}. \tag{8.37}$$

Substituting the constraints 8.34 to 8.37 into the model 8.33 yields (after some re-arrangement)

$$y_i^* = \beta_0 + \beta_2 z_{2i} + \beta_3 z_{3i} + \beta_{12} z_{12i} + \beta_{13} z_{13i} + \beta_{23} z_{23i} + v_i \tag{8.38}$$

where

$$y_i^* \equiv \ln q_i - \ln x_{1i}, \tag{8.39}$$
$$z_{ni} \equiv \ln x_{ni} - \ln x_{1i}, \tag{8.40}$$

and

$$z_{nmi} \equiv \ln x_{ni} \ln x_{mi} - 0.5\left(\ln x_{ni}\right)^2 - 0.5\left(\ln x_{mi}\right)^2. \tag{8.41}$$

This so-called *restricted model* is linear in the parameters and can be estimated by OLS. Following estimation, estimates of β_1, β_{11}, β_{22} and β_{33} can be computed by substituting the estimates of β_2, β_3, β_{12}, β_{13} and β_{23} into the right-hand sides of the constraints 8.34 to 8.37.

b) *If the regression model is nonlinear and/or one of the constraints is nonlinear* then the first-order conditions for the constrained optimisation problem are generally nonlinear and need to be solved numerically (i.e., using an iterative optimisation procedure). The substitution method described above can make the computations easier but is only feasible if each constraint can be solved uniquely for one of the parameters. An example of a nonlinear constraint that does *not* satisfy this property is $\beta_1^2 + \beta_2^2 = 1$.

To illustrate these procedures, Table 8.4 presents annotated SHAZAM output from estimating the constant returns to scale translog model discussed above. Results from estimating the unrestricted version of this model are presented in Table 8.2. A comparison of the results reported in these two tables reveals that imposing constant returns to scale has a reasonably large impact on the area and labour elasticities (i.e., coefficients of LX1 and LX2). In the next section we consider methods for testing the validity of the constant returns to scale assumption.

Table 8.4 Constant Returns to Scale Translog Production Function

```
|_ols lq lx1-lx3 lx11-lx13 lx22-lx23 lx33 /restrict

REQUIRED MEMORY IS PAR=     156 CURRENT PAR=   28000
 OLS ESTIMATION
      344 OBSERVATIONS      DEPENDENT VARIABLE= LQ
...NOTE..SAMPLE RANGE SET TO:      1,     344
|_restrict lx1+lx2+lx3 = 1
|_restrict lx11+lx12+lx13 = 0
|_restrict lx12+lx22+lx23 = 0
|_restrict lx13+lx23+lx33 = 0
|_end
F TEST ON RESTRICTIONS=   3.5550     WITH    4 AND  334 DF  P-VALUE= 0.00740

 R-SQUARE =    0.8654     R-SQUARE ADJUSTED =    0.8634
VARIANCE OF THE ESTIMATE-SIGMA**2 =  0.10497
STANDARD ERROR OF THE ESTIMATE-SIGMA =  0.32398
SUM OF SQUARED ERRORS-SSE=   35.478
MEAN OF DEPENDENT VARIABLE = -0.32631
LOG OF THE LIKELIHOOD FUNCTION = -97.3784

< … snip … >

VARIABLE     ESTIMATED   STANDARD   T-RATIO          PARTIAL STANDARDIZED ELASTICITY
  NAME      COEFFICIENT    ERROR     338 DF    P-VALUE CORR.  COEFFICIENT  AT MEANS

LX1        0.46874     0.74940E-01   6.2549     0.0000 0.3221  0.42791      0.39052
LX2        0.29324     0.70605E-01   4.1532     0.0000 0.2204  0.26612      0.24914
LX3        0.23802     0.47033E-01   5.0607     0.0000 0.2654  0.26359      0.29750
LX11      -0.83895     0.22027      -3.8087     0.0002-0.2029 -0.52182      0.91571
LX12       0.57831     0.21535       2.6855     0.0076 0.1445  0.66109      -1.1777
LX13       0.26064     0.13410       1.9436     0.0528 0.1051  0.35142     -0.62342
LX22      -0.34242     0.27525      -1.2440     0.2143-0.0675 -0.19653      0.37135
LX23      -0.23589     0.13428      -1.7567     0.0799-0.0951 -0.31187      0.57023
LX33      -0.24749E-01 0.95303E-01 -0.25969     0.7953-0.0141 -0.22395E-01  0.41935E-01
CONSTANT   0.11522E-01 0.21446E-01  0.53723     0.5915 0.0292   0.0000     -0.35309E-01

< … snip … >

|_genr ystar = lq - lx1
|_genr z2 = lx2 - lx1
|_genr z3 = lx3 - lx1
|_genr z12 = lx12 - lx11 - lx22
|_genr z13 = lx13 - lx11 - lx33
|_genr z23 = lx23 - lx22 - lx33

|_ols ystar z2 z3 z12 z13 z23

REQUIRED MEMORY IS PAR=     161 CURRENT PAR=   28000
 OLS ESTIMATION
      344 OBSERVATIONS      DEPENDENT VARIABLE= YSTAR
...NOTE..SAMPLE RANGE SET TO:      1,     344

 R-SQUARE =    0.3232     R-SQUARE ADJUSTED =    0.3132
VARIANCE OF THE ESTIMATE-SIGMA**2 =  0.10497
STANDARD ERROR OF THE ESTIMATE-SIGMA =  0.32398
SUM OF SQUARED ERRORS-SSE=   35.478
MEAN OF DEPENDENT VARIABLE = -0.54453E-01
LOG OF THE LIKELIHOOD FUNCTION = -97.3784

< … snip … >

VARIABLE     ESTIMATED   STANDARD   T-RATIO          PARTIAL STANDARDIZED ELASTICITY
  NAME      COEFFICIENT    ERROR     338 DF    P-VALUE CORR.  COEFFICIENT  AT MEANS

Z2         0.29324     0.70605E-01   4.1532     0.0000 0.2204  0.22672      0.28975E-01
Z3         0.23802     0.47033E-01   5.0607     0.0000 0.2654  0.29787      0.59446
Z12        0.57831     0.21535       2.6855     0.0076 0.1445  0.13106      0.48383
Z13        0.26064     0.13410       1.9436     0.0528 0.1051  0.17088      0.61537
Z23       -0.23589     0.13428      -1.7567     0.0799-0.0951 -0.13534     -0.51104
CONSTANT   0.11522E-01 0.21445E-01  0.5372     0.591 0.029     0.0000     -0.2116
|_test 1 - z2 - z3
TEST VALUE =  0.46874     STD. ERROR OF TEST VALUE  0.74940E-01
T STATISTIC =   6.2548522     WITH  338 D.F.    P-VALUE= 0.00000
F STATISTIC =   39.123176     WITH    1 AND  338 D.F.  P-VALUE= 0.00000
WALD CHI-SQUARE STATISTIC =   39.123176     WITH    1 D.F.  P-VALUE= 0.00000
UPPER BOUND ON P-VALUE BY CHEBYCHEV INEQUALITY = 0.02556
|_test - z12 - z13
TEST VALUE = -0.83895     STD. ERROR OF TEST VALUE  0.22027

< … snip … >
```

Using the RESTRICT option to impose the homogeneity restrictions 8.29 to 8.32.

An alternative to using the RESTRICT option is to generate variables using 8.39 to 8.41 and then estimate the restricted model 8.38 using OLS.

Both sets of commands yield the same RLS estimates.

Some parameters were substituted out of the model. We can recover estimates of these parameters and their standard errors using TEST commands.

8.5 Hypothesis Testing

If the errors are normally distributed, or if our sample size is large, we can test hypotheses concerning a single coefficient using a t-test. Let β_k denote the k-th element of the vector $\boldsymbol{\beta}$ and let c be a known constant. To test H_0: $\beta_k = c$ against H_1: $\beta_k \neq c$ we use the test statistic

$$t = \frac{b_k - c}{se(b_k)} \sim t(I - K) \tag{8.42}$$

where b_k is our estimator for β_k and $se(b_k)$ is the estimator for its standard error.[10] Thus, we reject H_0 at the $100\alpha\%$ level of significance if the absolute value of the test statistic is greater than the critical value $t_{1-\alpha/2}(I - K)$. If the alternative hypothesis is H_1: $\beta_k < c$ then we reject H_0 if the t-statistic is less than $t_\alpha(I - K)$, and if the alternative hypothesis is H_1: $\beta_k > c$ we reject H_0 if the t-statistic is greater than $t_{1-\alpha}(I - K)$. For example, we can use the results reported in Table 8.4 to test H_0: $\beta_2 = 0$ against H_1: $\beta_2 > 0$ at the $\alpha = 0.05$ level of significance. The t-statistic reported in Table 8.4 is computed as

$$t = \frac{b_2 - 0}{se(b_2)} = \frac{0.29324}{0.07061} = 4.153. \tag{8.43}$$

The test statistic is greater than the critical value $t_{0.95}(334) = 1.645$ so we reject H_0 and conclude that β_2 (the labour elasticity) is positive (at the 5% level of significance).

Sometimes we wish to conduct a joint test of several hypotheses concerning the coefficients. For example, we may be interested in testing the joint null hypothesis that all regression coefficients apart from the intercept term are zero (this is known as a *test of the significance of the regression*) or we may want to test whether a translog production function exhibits constant returns to scale (these constraints are discussed in Section 8.4). Several procedures are available for testing hypotheses of this form. All are underpinned by the idea that if the restrictions specified under the null hypothesis are true then our restricted and unrestricted estimates of $\boldsymbol{\beta}$ should be very close (because our restricted and unrestricted estimators are both consistent). Among other things, this implies that

1. The difference between the sums of squared residuals obtained from the restricted and unrestricted models should be close to zero. A measure of this closeness is the F-statistic

[10] We follow Abadir and Magnus (2002) and use the notation $x \sim t(\upsilon)$ to mean that x has a t-distribution with υ degrees of freedom. We also use the notation $t_\alpha(\upsilon)$ to denote the $100\alpha\%$ quantile of the distribution (i.e. the value that leaves an area of α in the left-hand-tail of the pdf). Similar notation is used for chi-squared, standard normal and F distributions.

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(I-K)} \sim F(J, I-K) \tag{8.44}$$

where SSE_R and SSE_U are the restricted and unrestricted sums of squared residuals and J is the number of restrictions. We reject H_0 at the $100\alpha\%$ level of significance if the F-statistic exceeds the critical value $F_{1-\alpha}(J, I-K)$.[11] To illustrate, we can use the sums of squared errors reported in Tables 8.2 and 8.4 to test the null hypothesis that the Philippine rice production technology exhibits constant returns to scale. The test statistic is

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(I-K)} = \frac{(35.478 - 34.029)/4}{34.029/(344-10)} = 3.56 \tag{8.45}$$

(this value is also reported in Table 8.4). The critical value at the 5% level of significance is $F_{0.95}(4, 334) = 2.37$ so we reject the null hypothesis and conclude that the technology does not exhibit constant returns to scale.

2. The value of the log-likelihood function evaluated at the unrestricted estimates should be close to the value of the log-likelihood function evaluated at the restricted estimates. A measure of this closeness is the likelihood ratio statistic:

$$LR = -2\left[\ln L_R - \ln L_U\right] \sim \chi^2(J) \tag{8.46}$$

where $\ln L_R$ and $\ln L_U$ denote the maximised values of the restricted and unrestricted log-likelihood functions and J is still the number of restrictions. We reject H_0 at the $100\alpha\%$ level of significance if the LR statistic exceeds the critical value $\chi^2_{1-\alpha}(J)$. For example, to test the null hypothesis of constant returns to scale at the $\alpha = 0.05$ level of significance we use the results reported in Tables 8.2 and 8.4 to compute

$$LR = -2\left[-97.3784 + 90.2070\right] = 14.34. \tag{8.47}$$

The critical value is $\chi^2_{0.95}(4) = 9.488$ so we again reject the null hypothesis and conclude that the technology does not exhibit constant returns to scale.

The F and LR tests are simple to construct but require estimation of both the restricted and unrestricted models. This may be problematic if one or both models are complicated and the availability of computer time is limited. Alternative testing procedures that require estimation of only one model are the *Wald (W)* and

[11] The t- and F-tests are equivalent in the special case where there is only one restriction involving only one parameter. The value of the F-statistic is the square of the t-statistic and the F-critical value is the square of the t-critical value. Thus, the numerical values of the test statistics and critical values differ but the statistical conclusions are the same.

Lagrange Multiplier (LM) tests. The Wald statistic measures how well the restrictions are satisfied when evaluated at the unrestricted estimates (so it only requires estimation of the unrestricted model). The *LM* statistic looks at the first-order conditions for a maximum of the log-likelihood function when evaluated at the restricted estimates (so it only requires estimation of the restricted model). For more details see Greene (2003).

All the tests discussed in this section are asymptotically justified (i.e., we can use them if the sample size is large). Moreover, the *F*, *LR*, *W* and *LM* tests are all asymptotically equivalent. Thus, if our sample is large enough, the four tests should yield the same results and we can simply choose a test procedure on the basis of computational convenience (which often depends on our computer software). However, matters are not so straightforward if our sample size is small. The *t*- and *F*-tests are justified in small samples provided the model and constraints are linear in the parameters and the errors are normally distributed. Unfortunately, even in this special case, we can't say much about the small-sample properties of the *LR*, *W* and *LM* statistics except that they satisfy the inequality $W \geq LR \geq LM$.

8.6 Systems Estimation

We often want to estimate the parameters of several economic relationships, each of which can be written in the form of the linear regression model 8.12. For example, consider the single-output three-input translog cost function

$$\ln c_i = \beta_0 + \sum_{n=1}^{3} \beta_n \ln w_{ni} + \beta_q \ln q_i + 0.5 \sum_{n=1}^{3} \sum_{m=1}^{3} \beta_{nm} \ln w_{ni} \ln w_{mi} + \sum_{n=1}^{3} \beta_{qn} \ln q_t \ln w_{ni} + \beta_{qq} \left(\ln q_i\right)^2 + v_i \tag{8.48}$$

where c_i is the i-th observation on cost, the w_{ni} are input prices and q_i is output. We can write this model in the form of 8.12 and obtain estimates of the parameters using OLS or ML. However, we can also use Shephard's Lemma (see Chapter 2) to derive the cost-share equations

$$s_{nt} = \beta_n + \sum_{m=1}^{3} \beta_{nm} \ln w_{mi} + \beta_{qn} \ln q_i + v_{ni} \qquad \text{for } n = 1, 2, 3, \tag{8.49}$$

where s_{ni} is the n-th cost share and the v_{ni}s are random errors that are likely to be correlated with each other and with the errors in 8.48. The correlations between these errors provide information that can be used in the estimation process to obtain more efficient estimates. In the present context, this means estimating a system of $N = 3$ equations comprising the cost function and, for reasons given below, only $N - 1 = 2$ of the cost-share equations.

Systems of equations such as this, where the errors in the equations are correlated, are known as *seemingly unrelated regression (SUR)* models. Again, the parameters of SUR models can be estimated using least squares or maximum likelihood techniques.

One method for obtaining least squares estimates is to estimate each equation separately by OLS and use the residuals to estimate the variances and covariances of the errors. We then minimise a generalised sum of squares function that is similar to 8.19 except it involves these estimated variances and covariances. Solving this optimisation problem yields the *estimated generalised least squares (EGLS)* estimator of β.

Because EGLS gives a more efficient estimator than OLS (provided the correlations among the errors are sufficiently large), it is possible to use the residuals to re-estimate the variances and covariances of the errors, then minimise a generalised sum of squares function and obtain a new estimate of β. Such estimates are sometimes known as *iterative EGLS* estimates. Interestingly, if we repeat this updating process to the point where successive iterations make little or no difference to our coefficient estimates, our estimates will be identical to ML estimates obtained under the assumption that the error terms are normally distributed. This iterative procedure is automated in most econometrics software packages.

Before considering an empirical example, there are several important practical issues that need mention:

- if the dependent variables in a set of SUR equations sum to the value of a variable appearing on the right-hand sides of those equations (eg., when share equations include an intercept term) the covariance matrix of the errors is usually singular. In this case, unless one equation is deleted from the system the estimation breaks down. If we estimate the SUR model by iterative EGLS then it is immaterial which equation is deleted.
- EGLS estimation of the SUR model is identical to single-equation OLS estimation if exactly the same explanatory variables appear in each equation and the parameters are not required to satisfy any cross-equation restrictions.
- EGLS is also identical to single-equation OLS if the covariances between the errors in the different equations are all zero and the parameters are not required to satisfy any cross-equation restrictions. We can test the null hypothesis that all covariances among the errors in different equations are zero using an LR test.[12]

To illustrate these ideas, Table 8.5 presents SHAZAM output from the estimation of the cost and cost-share system 8.48 and 8.49. Again, the data is the Philippine

[12] An alternative test based on OLS residuals is an LM test suggested by Breusch and Pagan (1980). For details see Greene (2003, p.350-351).

rice data discussed in Chapter 5. The variables have been scaled to have unit means, so the first-order coefficients can be interpreted as estimated cost shares evaluated at the variable means. The estimates reported in Table 8.5 are obtained by deleting the labour share equation from the SUR system. Note that we reject the null hypothesis that all the cross-equation error covariances are zero at any level of significance.

8.7 Inequality Constraints

Several of the regularity conditions discussed in Chapter 2 can be written in the form of inequality constraints involving the parameters. For example, a Cobb-Douglas production function is non-decreasing in inputs if each first-order coefficient is non-negative, and a quadratic cost function is concave in prices if the (symmetric) matrix of second-order price coefficients is negative semi-definite.

A form of the substitution method discussed in Section 8.4 can be used to impose these types of constraints. For example, non-negativity constraints of the form $\beta \geq 0$ can be enforced by replacing β in our model with either α^2 or e^{α} (both of which are non-negative irrespective of the value of α). Constraining a symmetric matrix to be negative semi-definite is also reasonably straightforward using the Cholesky decomposition method suggested by Lau (1978) and illustrated by Jorgenson and Fraumeni (1981). For example, to impose the condition that the symmetric matrix

$$\mathbf{B} = \begin{bmatrix} \beta_{11} & \beta_{12} & \beta_{13} \\ \beta_{12} & \beta_{22} & \beta_{23} \\ \beta_{13} & \beta_{23} & \beta_{33} \end{bmatrix} \tag{8.50}$$

is negative semi-definite we first replace each element of **B** by the corresponding element of

$$\mathbf{LDL}' = \begin{bmatrix} 1 & 0 & 0 \\ \lambda_{12} & 1 & 0 \\ \lambda_{13} & \lambda_{23} & 1 \end{bmatrix} \begin{bmatrix} \delta_1 & 0 & 0 \\ 0 & \delta_2 & 0 \\ 0 & 0 & \delta_3 \end{bmatrix} \begin{bmatrix} 1 & \lambda_{12} & \lambda_{13} \\ 0 & 1 & \lambda_{23} \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \delta_1 & \lambda_{12}\delta_1 & \lambda_{13}\delta_1 \\ \lambda_{12}\delta_1 & \lambda_{12}\lambda_{12}\delta_1 + \delta_2 & \lambda_{12}\lambda_{13}\delta_1 + \lambda_{23}\delta_2 \\ \lambda_{13}\delta_1 & \lambda_{12}\lambda_{13}\delta_1 + \lambda_{23}\delta_2 & \lambda_{13}\lambda_{13}\delta_1 + \lambda_{23}\lambda_{23}\delta_2 + \delta_3 \end{bmatrix}. \tag{8.51}$$

We then constrain the δ-parameters to be non-positive.[13] If the model isn't nonlinear already, making these substitutions and imposing non-negativity constraints will result in a nonlinear model that can be estimated using the techniques discussed in Section 8.3.3.

[13] These parameters are known as *Cholesky values*.

Table 8.5 Systems Estimation of a Translog Cost Function

```
_system 3/ restrict noconstant iter=100
_ols lc constant lw1-lw3 lq lw11-lw13 lw22-lw23 lw33 lqw1 lqw2 lqw3 lqq
_ols s1 constant lw1-lw3 lq
_ols s2 constant lw1-lw3 lq
_restrict lw1:1 + lw2:1 + lw3:1  = 1
_restrict lw11:1 + lw12:1 + lw13:1 = 0
_restrict lw12:1 + lw22:1 + lw23:1 = 0
_restrict lw13:1 + lw23:1 + lw33:1 = 0
_restrict lqw1:1 + lqw2:1 + lqw3:1 = 0
_restrict constant:2 - lw1:1 = 0
_restrict lw1:2 - lw11:1 = 0
_restrict lw2:2 - lw12:1 = 0
_restrict lw3:2 - lw13:1 = 0
_restrict lq:2 - lqw1:1 = 0
_restrict constant:3 - lw2:1 = 0
_restrict lw1:3 - lw12:1 = 0
_restrict lw2:3 - lw22:1 = 0
_restrict lw3:3 - lw23:1 = 0
_restrict lq:3 - lqw2:1 = 0
_end
MULTIVARIATE REGRESSION--    3 EQUATIONS
  25 RIGHT-HAND SIDE VARIABLES IN SYSTEM
MAX ITERATIONS =      100          CONVERGENCE TOLERANCE =    0.10000E-02
      344 OBSERVATIONS
IR OPTION IN EFFECT - ITERATIVE RESTRICTIONS

< … snip … >

ITERATION    17 SIGMA INVERSE
  25.030
  109.67        928.18
  47.343        554.48        743.89

ITERATION     17 COEFFICIENTS
   9.9800       0.50737       0.36028       0.13235       0.86210       0.12618
 -0.12370      -0.24808E-02   0.13444      -0.10743E-01   0.13223E-01   0.23700E-01
 -0.19704E-01  -0.39961E-02   0.33073E-01   0.50737       0.12618      -0.12370
 -0.24808E-02   0.23700E-01   0.36028      -0.12370       0.13444      -0.10743E-01
 -0.19704E-01

ITERATION    17 SIGMA
 0.88726E-01
-0.12818E-01  0.37940E-02
 0.39078E-02 -0.20122E-02  0.25954E-02
LOG OF DETERMINANT OF SIGMA=   -15.278
LOG OF LIKELIHOOD FUNCTION =    1163.48

LIKELIHOOD RATIO TEST OF DIAGONAL COVARIANCE MATRIX =    851.36
CHI-SQUARE WITH    3 D.F.     P-VALUE= 0.00000

< … snip … >

 EQUATION  1 OF  3 EQUATIONS
DEPENDENT VARIABLE = LC                    344 OBSERVATIONS

< … snip … >

VARIABLE     ESTIMATED    STANDARD    T-RATIO          PARTIAL STANDARDIZED ELASTICITY
  NAME      COEFFICIENT     ERROR      344 DF    P-VALUE CORR.  COEFFICIENT  AT MEANS

CONSTANT    9.9800      0.17482E-01   570.88     0.0000 0.9995   0.0000        1.0377
LW1        0.50737      0.35096E-02   144.57     0.0000 0.9919   0.39463      -0.10831E-01
LW2        0.36028      0.29276E-02   123.06     0.0000 0.9888   0.15023      -0.24205E-02
LW3        0.13235      0.27980E-02   47.302     0.0000 0.9310   0.31958E-01  -0.31633E-03
LQ         0.86210      0.21964E-01   39.250     0.0000 0.9041   0.86804      -0.29250E-01
LW11       0.12618      0.36130E-02   34.924     0.0000 0.8832   0.68933E-01   0.32752E-02
LW12      -0.12370      0.36297E-02  -34.079     0.0000-0.8783  -0.38119E-01  -0.61809E-03
LW13      -0.24808E-02  0.29329E-02  -0.84586    0.3982-0.0456  -0.55103E-03   0.13277E-05
LW22       0.13444      0.61521E-02   21.853     0.0000 0.7624   0.15918E-01   0.94748E-03
LW23      -0.10743E-01  0.50285E-02  -2.1364     0.0334-0.1144  -0.10713E-02  -0.85809E-05
LW33       0.13223E-01  0.53788E-02   2.4584     0.0144 0.1314   0.62215E-03   0.30650E-04
LQW1       0.23700E-01  0.37934E-02   6.2477     0.0000 0.3192   0.29102E-01   0.68409E-03
LQW2      -0.19704E-01  0.32334E-02  -6.0937     0.0000-0.3121  -0.76643E-02  -0.10149E-03
LQW3      -0.39961E-02  0.30681E-02  -1.3025     0.1936-0.0701  -0.10513E-02  -0.24787E-05
LQQ        0.33073E-01  0.98472E-02   3.3587     0.0009 0.1782   0.57694E-01   0.30005E-02
```

These constraints ensure the cost function is homogeneous of degree one in input prices, as required by economic theory – see Chapter 2.

These are cross-equation restrictions that ensure, for example, that the coefficient of LW1 in the cost function is the same as the intercept term in the first share equation.

The procedure has converged to the ML estimates after 17 iterations. These are estimates of the elements of the error covariance matrix.

The LR test of the null hypothesis that the cross-equation error covariances are all zero.

Perhaps because of its relative simplicity, the substitution approach is widely used in practice to impose economic regularity conditions including monotonicity (i.e., the non-increasing and non-decreasing properties of functions) and curvature (eg., the concavity and convexity properties). However, we note two points:

1. It is not always possible to devise substitutions that will ensure economic regularity conditions are satisfied. For example, there are no convenient substitutions that ensure translog functions are *quasi*-concave at all possible data points (i.e., globally).
2. Imposing global curvature constraints often destroys the flexibility properties of second-order flexible functional forms (Diewert and Wales, 1987). One solution is to impose these constraints at only a handful of data points in the region of the point of approximation (i.e., locally). Unfortunately, the substitution method doesn't lend itself easily to the imposition of local curvature constraints, except in the trivial case where they are imposed at a single point.[14]

To illustrate some of these ideas, Table 8.6 presents annotated SHAZAM output obtained from imposing linear homogeneity and global concavity on the translog cost function model that is estimated in Section 8.6. This involves replacing the second-order parameters in 8.48 and 8.49 with[15]

$$\begin{aligned} \beta_{11} &= -\alpha_1^2 & \beta_{22} &= -(1+\lambda_{13})^2\alpha_1^2 - \alpha_2^2 \\ \beta_{12} &= (1+\lambda_{13})\alpha_1^2 & \beta_{23} &= (1+\lambda_{13})\lambda_{13}\alpha_1^2 + \alpha_2^2 \\ \beta_{13} &= -\lambda_{13}\alpha_1^2 & \beta_{33} &= -\lambda_{13}^2\alpha_1^2 - \alpha_2^2 \end{aligned} \tag{8.52}$$

From Table 8.6, we see that these constraints have forced the second-order price coefficients to zero. This illustrates how imposing global concavity can destroy the flexibility properties of second-order flexible functional forms.

Many researchers choose to avoid these difficulties by simply estimating an unrestricted model and assessing the severity of any constraint violations. For example, it is common practice to use the second-order derivatives of an estimated function to check a concavity or convexity condition at every point in the data set and then report the proportion of times the condition is violated[16]. More recently, Terrell (1996) has shown how to deal with the problem using a Bayesian approach.

---

[14] Ryan and Wales (2000) use the substitution method to ensure estimated translog and Generalised Leontief cost functions satisfy concavity at a single point.

[15] The matrix of second-order parameters is given by 8.50. We can ensure global concavity by constraining this matrix to be negative semi-definite. We do this by replacing the δ-paarameters in 8.51 with $-\alpha_i^2$. To ensure homogeneity of the cost function, we set $\delta_3 = 0$, $\lambda_{23} = -1$ and $\lambda_{12} = -(1 + \lambda_{13})$.

[16] Criteria for checking concavity and convexity can be found in Chiang (1984).

Table 8.6 Imposing Global Concavity on a Translog Cost Function

```
|_nl 3/ ncoef = 11
...NOTE..SAMPLE RANGE SET TO:      1,    344
|_eq lc = lw1 + b0 +  b2*(lw2 - lw1) +  b3*(lw3 - lw1) + bq*lq &
|       -  0.5*(a1**2)*lw1*lw1 + (1+lam13)*(a1**2)*lw1*lw2 - lam13*(a1**2)*lw1*lw3 &
|       -  0.5*((1+lam13)*(1+lam13)*(a1**2) + a2**2)*lw2*lw2 &
|                + ((1+lam13)*lam13*(a1**2) + a2**2)*lw2*lw3 &
|       -  0.5*(lam13*lam13*(a1**2) + a2**2)*lw3*lw3 &
|       + bq2*lq*(lw2 - lw1) + bq3*lq*(lw3 - lw1) + bqq*lq*lq
|_eq s1 = b1 - (a1**2)*lw1 + (1+lam13)*(a1**2)*lw2 - lam13*(a1**2)*lw3 - (bq2 + bq3)*lq
|_eq s2 = b2 + (1+lam13)*(a1**2)*lw1 -((1+lam13)*(1+lam13)*(a1**2) + a2**2)*lw2 &
          + ((1+lam13)*lam13*(a1**2) + a2**2)*lw3 + bq2*lq
|_end

< … snip … >

FINAL STATISTICS :

TIME =       0.765 SEC.   ITER. NO.    41   FUNCT. EVALUATIONS    72
LOG-LIKELIHOOD FUNCTION=     922.2788
COEFFICIENTS
   9.975128      0.3578347      0.3132100      0.9031072      0.5600096E-07
  -0.6102070     0.3113715E-07 -0.5517581E-01 -0.4355096E-02  0.2000187E-01
   0.5127308
GRADIENT
   0.1633178E-02  0.1639502E-01 -0.2381175E-03 -0.1845679E-02 -0.1614387E-02
  -0.1209324E-10  0.2051145E-02  0.9423735E-03  0.1157805E-01  0.1411328E-02
   0.1375700E-01

SIGMA MATRIX
 0.53134E-01
-0.13761E-01  0.15280E-01
 0.73620E-02 -0.11921E-01  0.11092E-01
GTRANSPOSE*INVERSE(H)*G  STATISTIC  -  =  0.53760E-10

      COEFFICIENT     ST. ERROR    T-RATIO

B0        9.9751      0.13876E-01  718.87
B2       0.35783      0.61554E-02  58.133
B3       0.31321      0.27445E-01  11.412
BQ       0.90311      0.20339E-01  44.402
A1       0.56001E-07  0.15902E-01 0.35215E-05
LAM13   -0.61021      0.14125E-01 -43.200
A2       0.31137E-07  0.22983E-01 0.13548E-05
BQ2     -0.55176E-01  0.61699E-02 -8.9428
BQ3     -0.43551E-02  0.30210E-02 -1.4416
BQQ      0.20002E-01  0.96849E-02  2.0653
B1       0.51273      0.71801E-02  71.410
|_end
|_test - (a1**2)
TEST VALUE = -0.31361E-14 STD. ERROR OF TEST VALUE  0.17811E-08
ASYMPTOTIC NORMAL STATISTIC = -0.17607674E-05  P-VALUE= 1.00000
WALD CHI-SQUARE STATISTIC =  0.31003019E-11 WITH    1 D.F.  P-VALUE= 1.00000
UPPER BOUND ON P-VALUE BY CHEBYCHEV INEQUALITY = 1.00000
|_test (1+lam13)*(a1**2)
TEST VALUE =  0.12224E-14 STD. ERROR OF TEST VALUE  0.69426E-09
ASYMPTOTIC NORMAL STATISTIC =  0.17607674E-05  P-VALUE= 1.00000
WALD CHI-SQUARE STATISTIC =  0.31003020E-11 WITH    1 D.F.  P-VALUE= 1.00000
UPPER BOUND ON P-VALUE BY CHEBYCHEV INEQUALITY = 1.00000
|_test - lam13*(a1**2)
TEST VALUE =  0.19137E-14 STD. ERROR OF TEST VALUE  0.10868E-08
ASYMPTOTIC NORMAL STATISTIC =  0.17607674E-05  P-VALUE= 1.00000
WALD CHI-SQUARE STATISTIC =  0.31003018E-11 WITH    1 D.F.  P-VALUE= 1.00000
UPPER BOUND ON P-VALUE BY CHEBYCHEV INEQUALITY = 1.00000
|_test - (1+lam13)*(1+lam13)*(a1**2) - a2**2
TEST VALUE = -0.14460E-14 STD. ERROR OF TEST VALUE  0.14558E-08
ASYMPTOTIC NORMAL STATISTIC = -0.99328365E-06  P-VALUE= 1.00000
WALD CHI-SQUARE STATISTIC =  0.98661242E-12 WITH    1 D.F.  P-VALUE= 1.00000
UPPER BOUND ON P-VALUE BY CHEBYCHEV INEQUALITY = 1.00000
|_test (1+lam13)*lam13*(a1**2) + a2**2
TEST VALUE =  0.22358E-15 STD. ERROR OF TEST VALUE  0.14939E-08
ASYMPTOTIC NORMAL STATISTIC =  0.14966495E-06  P-VALUE= 1.00000
WALD CHI-SQUARE STATISTIC =  0.22399598E-13 WITH    1 D.F.  P-VALUE= 1.00000
UPPER BOUND ON P-VALUE BY CHEBYCHEV INEQUALITY = 1.00000
|_test - lam13*lam13*(a1**2) - a2**2
TEST VALUE = -0.21373E-14 STD. ERROR OF TEST VALUE  0.15756E-08
ASYMPTOTIC NORMAL STATISTIC = -0.13564912E-05  P-VALUE= 1.00000
WALD CHI-SQUARE STATISTIC =  0.18400684E-11 WITH    1 D.F.  P-VALUE= 1.00000
UPPER BOUND ON P-VALUE BY CHEBYCHEV INEQUALITY = 1.00000
```

To impose both homogeneity and concavity we replace β_{12} with $\beta_{12} = (1+\lambda_{13})\alpha_1^2$ etc.

We can recover estimates of the second-order coefficients and their estimated standard errors using TEST commands. For example, this is the estimate of β_{12}.

8.8 The Bayesian Approach*

Until now, we have been employing the sampling theory approach to statistical inference. Some of the characteristics of this approach are:

1. Estimators are chosen on the basis of their long-run performance in (hypothetical) repeated samples. For example, in a repeated sampling context the OLS estimator 8.20 is the BLUE of $\boldsymbol{\beta}$ (see Section 8.3.1). Unfortunately, because we rarely collect more than one sample, we can rarely say anything about the accuracy of the numerical estimates obtained by plugging our data into formulas such as 8.20.
2. Similarly, we cannot make probability statements about unknown parameters, hypotheses or models. Even if a 95% confidence interval for β_k extends from 3.4 to 4.7, we cannot say that $3.4 < \beta_k < 4.7$ with probability 0.95. This is because β_k either lies in this interval or it doesn't, so the probability it lies in the interval is either 1 or 0, not 0.95.[17]
3. There are no convenient methods for incorporating some types of non-sample information into the estimation process (see Section 8.7).
4. It is difficult to obtain exact finite-sample results for certain estimation problems (see Section 8.3.3).

In this section, we consider the Bayesian approach to inference. This alternative approach to inference has the following characteristics:

1. Estimators are chosen based on their ability to minimise the loss associated with an estimation error.
2. Results are usually presented in terms of probability density functions (pdfs). Thus, it is possible and convenient to make probability statements about unknown parameters, hypotheses and models.
3. There is a formal mechanism for incorporating non-sample information into the estimation process.
4. Exact finite-sample results can be obtained for most estimation problems.

This section is concerned only with the application of Bayesian methods to the classical linear regression model with normally distributed errors. Bayesian estimation of stochastic frontier models is discussed in Chapter 10. More details concerning Bayesian estimation of these and other models are in Koop (2003).

---

* This section contains advanced material that could be optional in an introductory course.

[17] The interval from 3.4 to 4.7 does have a sampling theory interpretation. Specifically, if we collected many data sets, and each time we used the 95% confidence interval estimator to compute a confidence interval, approximately 95% of these intervals would contain β_k. The interval from 3.4 to 4.7 is one such interval.

8.8.1 Bayes' Theorem

The unknown parameters in the classical linear regression model are $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$ and σ. In the Bayesian approach to inference, we summarise pre-sample information about these parameters (eg., information from economic theory) in the form of a *prior* pdf, denoted $p(\boldsymbol{\beta}, \sigma)$. Sample information (i.e., information contained in the data) is summarised in the form of the familiar *likelihood function*, $L(\mathbf{y} \mid \boldsymbol{\beta}, \sigma)$. These two types of information are then combined using Bayes' Theorem:[18]

$$p(\boldsymbol{\beta}, \sigma \mid \mathbf{y}) \propto L(\mathbf{y} \mid \boldsymbol{\beta}, \sigma) p(\boldsymbol{\beta}, \sigma) \tag{8.53}$$

where $p(\boldsymbol{\beta}, \sigma \mid \mathbf{y})$ is the *posterior* pdf and $\propto$ denotes "is proportional to". In words, the posterior pdf is proportional to the likelihood function times the prior pdf. The posterior pdf underpins all types of Bayesian inference, including point and interval estimation, evaluation of hypotheses, and prediction.

8.8.2 Specifying a Prior Pdf

Prior pdfs are often classified as *non-informative* or *informative*. As the name suggests, a non-informative prior conveys ignorance about the parameters. In the case of the classical linear model with normal errors it is common practice to use the non-informative prior

$$p(\boldsymbol{\beta}, \sigma) \propto \frac{1}{\sigma} \tag{8.54}$$

which follows from the assumption that $\ln \sigma$ and the elements of $\boldsymbol{\beta}$ are all independently distributed and can take any values with equal probability.[19] In contrast, an informative prior conveys some information about at least one parameter. For example, if economic theory states that the elements of $\boldsymbol{\beta}$ satisfy monotonicity constraints, we could use

$$p(\boldsymbol{\beta}, \sigma) \propto \frac{I(\boldsymbol{\beta})}{\sigma} \tag{8.55}$$

where $I(\boldsymbol{\beta})$ is an indicator function that takes the value 1 if $\boldsymbol{\beta}$ satisfies the constraints and takes the value 0 otherwise. This type of prior pdf provides a formal mechanism for dealing with inequality constraints of the type discussed in Section 8.7.

---

[18] Technically Bayes' Theorem states that $p(\boldsymbol{\beta}, \sigma \mid \mathbf{y}) = L(\mathbf{y} \mid \boldsymbol{\beta}, \sigma) p(\boldsymbol{\beta}, \sigma) / p(\mathbf{y})$ where $p(\mathbf{y})$ is the *marginal likelihood*. However, after $\mathbf{y}$ has been observed the marginal likelihood is a constant so we often write Bayes' Theorem in the form 8.53. Bayes' Theorem is simply a rearrangement of the formula for conditional probability found in most elementary statistics textbooks.

[19] This doesn't necessarily mean that functions of these parameters can take any values with equal probability. For example, non-informative priors on the parameters of some stochastic frontier models can convey strong information about technical efficiency effects.

It is worth noting that the pdfs 8.54 and 8.55 are *improper* in the sense that they do not integrate to 1 (i.e., the area under the pdf is not 1). This does not present a problem provided the application of Bayes' Theorem yields a posterior pdf which is proper. Fortunately, it can be shown that combining either of the priors 8.54 or 8.55 with the likelihood function for the classical linear model with normal errors yields a proper posterior pdf.

8.8.3 Deriving Posterior Pdfs

According to Bayes' Theorem, the joint posterior pdf for $\boldsymbol{\beta}$ and σ is proportional to the likelihood function times a prior chosen by the researcher. For example, the likelihood function for the classical linear model with normal errors is given by 8.23, and combining this likelihood function with the noninformative prior 8.54 yields the posterior pdf

$$p(\boldsymbol{\beta},\sigma \mid \mathbf{y}) \propto \frac{1}{\sigma^{I+1}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{I}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2\right\}. \tag{8.56}$$

This joint posterior pdf summarises all our post-sample knowledge about $\boldsymbol{\beta}$ and σ^2.

Joint posterior pdfs are often functions of parameters that are of little interest (eg., σ). We can rid ourselves of these so-called *nuisance parameters* by integrating them out of the pdf. For example, it is possible (but difficult) to integrate 8.56 with respect to σ to obtain the *marginal posterior* pdf

$$p(\boldsymbol{\beta} \mid \mathbf{y}) \propto \left[\upsilon s^2 + (\boldsymbol{\beta} - \mathbf{b})'\left(\sum_{i=1}^{I}\mathbf{x}_i\mathbf{x}_i'\right)(\boldsymbol{\beta} - \mathbf{b})\right]^{-I/2} \tag{8.57}$$

where $\upsilon = I–K$ is a degrees of freedom parameter; $\mathbf{b}$ is the OLS estimator of $\boldsymbol{\beta}$; and

$$s^2 = \sum_{i=1}^{I}(y_i - \mathbf{x}_i'\mathbf{b})^2 \Big/ \upsilon \tag{8.58}$$

is the OLS estimator of σ^2. This pdf is in the form of a multivariate t-distribution with mean vector $\mathbf{b}$. With further integration, we can rid ourselves of $K–1$ elements of $\boldsymbol{\beta}$ and eventually obtain the marginal posterior pdf of the single element β_k. This pdf, denoted $p(\beta_k \mid \mathbf{y})$, takes the form of a univariate t-distribution with mean equal to the k-th element of $\mathbf{b}$.

Combining the likelihood function 8.23 with the informative prior 8.55 yields posterior pdfs that are identical to 8.56 and 8.57 except they are multiplied by the indicator function $I(\boldsymbol{\beta})$. Thus, the marginal posterior pdf $p(\boldsymbol{\beta} \mid \mathbf{y})$ will be in the form of a *truncated* multivariate t-distribution. The distribution is truncated because the

indicator function forces the posterior pdf to take the value zero if $\boldsymbol{\beta}$ fails to satisfy the constraints implied by economic theory. Unfortunately, the presence of the indicator function also means that further integration of $p(\boldsymbol{\beta} \mid \mathbf{y})$ with respect to the elements of $\boldsymbol{\beta}$ is likely to be analytically intractable – we usually need to evaluate such integrals using the simulation methods discussed in Section 8.9 below.

8.8.4 Point and Interval Estimation

A *loss function* expresses the loss associated with the error of estimation as a function of the estimator and the true value. If the loss function is a quadratic function then the optimal Bayesian point estimator (i.e., the estimator that minimises the expected loss) is the mean of the posterior pdf. For example, if we use the noninformative prior 8.54, the optimal Bayesian point estimator of $\boldsymbol{\beta}$ in the classical linear model with normal errors is the mean of the marginal posterior pdf 8.57 (i.e., the OLS estimator). If we are interested in the single element β_k, our point estimator is the mean of the marginal posterior pdf $p(\beta_k \mid \mathbf{y})$ (again, the OLS estimator of β_k). Thus, for example, the OLS coefficient estimates reported in Table 8.2 can be regarded as optimal Bayesian point estimates of the parameters of the translog production function model (under quadratic loss and using a noninformative prior).

We can use the marginal posterior pdf $p(\beta_k \mid \mathbf{y})$ to calculate the probability that β_k lies in a particular interval. More commonly, Bayesians like to specify the interval in which "most of the distribution lies". Such an interval, known as a *highest (posterior) density region (HDR)*, is the Bayesian counterpart to a confidence interval. A $(1-\alpha)\times 100\%$ HDR is the interval of shortest length that has area under the pdf equal to $(1-\alpha)$.

8.9 Simulation Methods

As we have seen, the Bayesian approach sometimes involves the evaluation of complex integrals. Historically, this has put Bayesian analysis beyond the reach of many applied economists. However, recent advances in computer technology and the theory of simulation now allow us to evaluate these integrals numerically.

To illustrate the idea, let $\boldsymbol{\theta}$ be a vector of unknown model parameters. Then almost anything a Bayesian would want to calculate can be written in the form

$$E\{h(\boldsymbol{\theta}) \mid \mathbf{y}\} = \int_{-\infty}^{\infty} h(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta} \tag{8.58}$$

where $h(\boldsymbol{\theta})$ is some function of $\boldsymbol{\theta}$ and $p(\boldsymbol{\theta} \mid \mathbf{y})$ is the pdf of $\boldsymbol{\theta}$ given $\mathbf{y}$. For example, if $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma)$ and we are interested in an optimal Bayesian point estimate of σ^2, we set $h(\boldsymbol{\theta}) = \sigma^2$. If we are interested in the probability that β_k is contained in an

interval we set $h(\boldsymbol{\theta})$ equal to an indicator function that takes the value 1 if β_k is contained in the interval and 0 otherwise.

Unfortunately, integrals of the form 8.58 are often analytically intractable. However, suppose $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots, \boldsymbol{\theta}^S$ is a random sample drawn from $p(\boldsymbol{\theta} \mid \mathbf{y})$. Then, provided S is large, we can estimate the integral 8.58 using the sample mean

$$\hat{h}(\boldsymbol{\theta}) = \frac{1}{S}\sum_{s=1}^{S} h(\boldsymbol{\theta}^s). \tag{8.59}$$

For example, to obtain a point estimate of σ^2 we can simply average sample observations on σ^2. To estimate the probability that β_k lies in an interval we calculate the proportion of sample draws on β_k that are contained in the interval.

In the remainder of this section, we discuss techniques for drawing random samples, or *simulating*, from a pdf. Because they involve random sampling these methods are also known as *Monte Carlo* methods. The simplest methods yield samples of independent observations. More sophisticated methods yield *chains* of correlated observations that satisfy the properties of Markov processes – these methods are known as *Markov Chain Monte Carlo (MCMC)* algorithms.

8.9.1 Methods for Drawing Independent Observations

If $p(\boldsymbol{\theta} \mid \mathbf{y})$ is of a recognisable or 'standard' form (eg., normal or gamma) then we can usually draw independent observations, $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots, \boldsymbol{\theta}^S$, using built-in functions in computer packages such as EViews and SHAZAM. For example, Table 8.7 shows how SHAZAM can be used to (slowly) draw a sample of observations from the multivariate t-distribution given by 8.57. The algorithm used in this table makes use of standard results for normal, chi-square and univariate and multivariate t-distributions available in Bratley, Fox and Schrage (1983, pp. 163-4). The empirical context is the translog production function model discussed in Section 8.3.1. For this model, the exact means of the marginal posterior pdfs are the coefficient estimates reported in Table 8.2. We could compute these numbers exactly by evaluating 8.58 analytically. Instead, Table 8.7 reports estimates computed using 8.59. Differences between the two sets of coefficient estimates are due to sampling error. Of course, we could reduce this sampling error by simply increasing the sample size.

Table 8.7 Bayesian Estimation of a Translog Production Function[a]

```
_set nooutput
_dim z 10 zero 10
_ols lq lx1-lx3 lx11-lx13 lx22-lx23 lx33 / coef = b cov = mlvar dn
_do ! = 1, 5000
_   matrix z = nor(10,1)
_   matrix chi = zero
_   do # = 1, 344
_      matrix chi = chi + nor(10,1)**2
_   endo
_   matrix t = z*sqrt(344)/sqrt(chi)
_   matrix beta = b + chol(mlvar)*t
_   write (beta.txt) beta / norewind
_endo
_rewind (beta.txt)
_smpl 1 5000
_read (beta.txt) blx1-blx3 blx11-blx13 blx22-blx23 blx33 bconst / eof
_set output
_stat blx1-blx3 blx11-blx13 blx22-blx23 blx33 bconst
NAME        N    MEAN          ST. DEV       VARIANCE      MINIMUM       MAXIMUM
BLX1        5000  0.57837      0.83767E-01 0.70170E-02  0.27653       0.88513
BLX2        5000  0.17972      0.79181E-01 0.62696E-02 -0.76344E-01  0.45567
BLX3        5000  0.21554      0.50054E-01 0.25054E-02  0.81643E-02  0.40427
BLX11       5000 -0.46655      0.24408     0.59574E-01  -1.2793      0.34428
BLX12       5000  0.68509      0.21458     0.46043E-01 -0.81114E-01   1.4504
BLX13       5000  0.73160E-01 0.14261     0.20336E-01 -0.46238       0.61134
BLX22       5000 -0.72801      0.29868     0.89212E-01  -1.8153      0.47210
BLX23       5000 -0.18843      0.13747     0.18899E-01 -0.59447       0.34340
BLX33       5000  0.25550E-01 0.96243E-01 0.92627E-02 -0.29574       0.35016
BCONST      5000  0.14368E-01 0.24366E-01 0.59370E-03 -0.65832E-01   0.11566
```

Obtain one draw and store it in a file.

We can estimate the mean of the marginal posterior pdf of β_2 using the mean of the sample observations on β_2.

When $p(\boldsymbol{\theta} \mid \mathbf{y})$ is of a non-standard form (eg., a truncated pdf), we can often draw independent observations using *accept-reject* methods. To do so, there must exist a pdf, $g(\boldsymbol{\theta} \mid \mathbf{y})$, which can be easily simulated from, and a constant, M, such that the inequality $p(\boldsymbol{\theta} \mid \mathbf{y}) \leq Mg(\boldsymbol{\theta} \mid \mathbf{y})$ is satisfied for all values of $\boldsymbol{\theta}$ where $p(\boldsymbol{\theta} \mid \mathbf{y}) > 0$. The first step is to draw a candidate $\boldsymbol{\theta}^*$ from $g(\boldsymbol{\theta} \mid \mathbf{y})$ and u from a standard uniform distribution. We then accept $\boldsymbol{\theta}^*$ as a draw from $p(\boldsymbol{\theta} \mid \mathbf{y})$ if and only if:

$$u \leq \frac{p(\boldsymbol{\theta}^* \mid \mathbf{y})}{Mg(\boldsymbol{\theta}^* \mid \mathbf{y})}. \tag{8.60}$$

In the special case where $p(\boldsymbol{\theta} \mid \mathbf{y})$ is a truncated distribution and $g(\boldsymbol{\theta} \mid \mathbf{y})$ is chosen to be the untruncated version of $p(\boldsymbol{\theta} \mid \mathbf{y})$, the accept-reject algorithm collapses to a very simple procedure – we simply draw $\boldsymbol{\theta}^*$ from $g(\boldsymbol{\theta} \mid \mathbf{y})$ and accept it if it lies in a region where the height of $p(\boldsymbol{\theta} \mid \mathbf{y})$ is non-zero (i.e., if it lies in the feasible region).

To motivate an empirical example, consider Table 8.7 and observe that the smallest sample observation on β_2 is less than zero. This means there is positive probability that the estimated function fails to satisfy monotonicity at the variable means.[20] This suggests that we should constrain all first-order coefficients to be non-negative. We can do this easily using the accept-reject algorithm – drawing from the truncated distribution is simply a matter of drawing from the untruncated

[20] Sampling theorists will also be unhappy with the results reported in Table 8.2 insofar as a lower bound on a 99% confidence interval for β_2 is less than zero.

distribution (as we did in Table 8.7) and throwing away those draws which fail to satisfy the constraints. Our parameter estimates are obtained by averaging over the draws that remain. We could do this by adding a few more lines of code to the SHAZAM code in Table 8.7. However, the procedure is already automated in SHAZAM – Table 8.8 shows how SHAZAM can be used to (quickly) draw a sample of observations from the posterior pdf and average over those draws that satisfy the constraints.

8.9.2 Markov Chain Monte Carlo Methods

Once we get beyond simple models, implementing the Bayesian approach often requires the use of an iterative MCMC algorithm. The two most popular algorithms are the *Metropolis-Hastings* algorithm and the *Gibbs Sampler*.

The Metropolis-Hastings (M-H) algorithm requires the specification of an arbitrary *transition* or *proposal* density, $q(\theta^* \mid \theta^s)$, that is easy to simulate from (eg., multivariate normal). We can then draw observations from the *target* density $p(\theta \mid \mathbf{y})$ using the following steps:

1. Choose a starting value θ^0 that is in the support[21] of θ and set $s = 0$.
2. Draw u from a standard uniform distribution and draw θ^* from $q(\theta^* \mid \theta^s)$.
3. Calculate $r = \min\left[\dfrac{p(\theta^* \mid \mathbf{y})q(\theta^s \mid \theta^*)}{p(\theta^s \mid \mathbf{y})q(\theta^* \mid \theta^s)}, 1\right]$.
4. If $u < r$ then set $\theta^{s+1} = \theta^*$; otherwise set $\theta^{s+1} = \theta^s$.
5. Set $s = s + 1$ and repeat from Step 2.

These steps ensure the candidate draw θ^* will be included in the MCMC sample with probability r. Many researchers choose a symmetric transition density such as a multivariate normal – in this case the ratio in Step 3 simplifies to $p(\theta^* \mid \mathbf{y})/p(\theta^s \mid \mathbf{y})$.

The idea behind the Metropolis-Hastings algorithm is illustrated in Figure 8.1 where we depict the problem of drawing an observation from a nonstandard univariate target density $p(\theta \mid \mathbf{y})$. In this figure, the symmetric proposal density, $q(\theta^* \mid \theta^s)$, is centred on θ^s (the last draw in the MCMC chain) and has been used to generate a candidate draw, θ^*. This draw is accepted into the chain with probability $r = p(\theta^* \mid \mathbf{y})/p(\theta^s \mid \mathbf{y})$. For the particular θ^* depicted in Figure 8.1, the value of r will be less than 0.5 because θ^* is further out in the tail of the target density than θ^s.

[21] This is the region where the height of the pdf is nonzero.

Table 8.8 Monotonicity-Constrained Translog Production Function

```
|_ols lq lx1-lx3 lx11-lx13 lx22-lx23 lx33

 < … snip … >

 |_bayes / nsamp = 5000
BAYESIAN (GEWEKE) INEQUALITY CONSTRAINED ESTIMATION
 |_restrict lx1.ge.0
 |_restrict lx2.ge.0
 |_restrict lx3.ge.0
 |_stop
NUMBER OF INEQUALITY RESTRICTIONS =    3
NUMBER OF COEFFICIENTS= 10
NUMBER OF REPLICATIONS=       5000
ANTITHETIC REPLICATIONS ALSO INCLUDED
DEGREES OF FREEDOM FOR T DISTRIBUTION =       334
ORIGINAL COEFFICIENT ESTIMATES
  0.57777        0.18002        0.21657       -0.46265
  0.69761E-01   -0.73097       -0.18651        0.27510E-01   0.13938E-01
     10000 REPLICATIONS        9877 SATISFIED
PROPORTION= 0.98770 NUMERICAL STANDARD ERROR OF PROPORTION= 0.00110
                    ASYMPTOTIC STANDARD ERROR OF PROPORTION= 0.00110
VARIABLE  AVERAGE       STDEV        VARIANCE     NUMERICAL SE
LX1       0.57575       0.84306E-01  0.71075E-02  0.84829E-03
LX2       0.18258       0.77499E-01  0.60061E-02  0.77980E-03
LX3       0.21625       0.49725E-01  0.24726E-02  0.50033E-03
LX11     -0.46174       0.24509      0.60069E-01  0.24661E-02
LX12      0.68544       0.21694      0.47063E-01  0.21829E-02
LX13      0.68728E-01   0.14604      0.21328E-01  0.14695E-02
LX22     -0.72982       0.30586      0.93552E-01  0.30776E-02
LX23     -0.18561       0.13949      0.19458E-01  0.14036E-02
LX33      0.27419E-01   0.97263E-01  0.94602E-02  0.97867E-03
CONSTANT  0.13918E-01   0.24763E-01  0.61323E-03  0.24917E-03
```

Monotonicity constraints

98.77% of draws from the untruncated distribution satisfy the constraints.

An estimate of β_2 is obtained by averaging over the draws that satisfy the constraints.

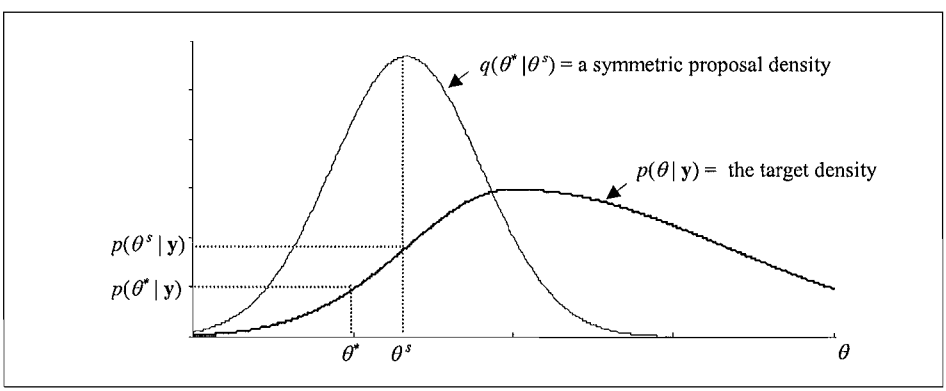

Figure 8.1 The Metropolis-Hastings Algorithm

The proportion of candidate draws that are accepted into a Metropolis-Hastings chain can be increased (decreased) by scaling down (up) the covariance matrix of the transition density. Roberts, Gelman and Gilks (1997) show that if the target and proposal densities are normal pdfs, the optimal acceptance rate (i.e., the one which minimises the autocorrelations across the sample values) is between approximately 0.25 and 0.45.

The Gibbs Sampler is an alternative MCMC algorithm that relies on our ability to partition $\boldsymbol{\theta}$ as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_P)$ where the $\boldsymbol{\theta}_p$s may be multi-dimensional and where it is possible to simulate from the conditional densities,

$$p(\boldsymbol{\theta}_p \mid \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{p-1}, \boldsymbol{\theta}_{p+1}, \ldots, \boldsymbol{\theta}_P, \mathbf{y}) \qquad \text{for } p = 1, \ldots, P.$$

We can then draw observations on $\boldsymbol{\theta}$ using the following steps:

1. Choose a starting value $\boldsymbol{\theta}^0$ that is in the support of $\boldsymbol{\theta}$ and set $s = 0$.
2. Draw $\boldsymbol{\theta}_1^{s+1}$ from $p(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2^s, \boldsymbol{\theta}_3^s, \ldots, \boldsymbol{\theta}_P^s, \mathbf{y})$.
3. Draw $\boldsymbol{\theta}_2^{s+1}$ from $p(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1^{s+1}, \boldsymbol{\theta}_3^s, \ldots, \boldsymbol{\theta}_P^s, \mathbf{y})$.

   : : : :

4. Draw $\boldsymbol{\theta}_P^{s+1}$ from $p(\boldsymbol{\theta}_P \mid \boldsymbol{\theta}_1^{s+1}, \boldsymbol{\theta}_2^{s+1}, \ldots, \boldsymbol{\theta}_{P-1}^{s+1}, \mathbf{y})$.
5. Set $s = s + 1$ and repeat from Step 2.

The Gibbs Sampler is particularly useful for problems involving latent variables (eg., tobit models and stochastic frontier models). How quickly the algorithm draws a sample depends on the methods used to sample from the conditional densities in each step. In complex models it may be necessary to sample from one or more of these densities using an accept-reject or other algorithm. For example, Terrell (1996) uses an accept-reject algorithm within the Gibbs to impose curvature constraints on a cost function, while O'Donnell and Coelli (2005) use an M-H algorithm within the Gibbs to impose curvature on an output distance function.[22]

More details on MCMC algorithms are available in Koop (2003). In practice, the large number of priors and likelihoods in econometrics makes it difficult to build a computer package that can be widely used for MCMC simulation. However, certain types of models can be estimated using packages such as BUGS (Bayesian Inference Using Gibbs Sampling), BACC (Bayesian Analysis, Computation and Communication) and BSFA (Bayesian Stochastic Frontier Analysis). Many researchers create their own programs using matrix languages such as GAUSS or MATLAB. It is also possible, but often less convenient, to do MCMC simulation using EViews or SHAZAM. An empirical example using SHAZAM and the Philippine rice data is presented in Section 10.7.

8.10 Conclusion

It is possible to estimate the important economic characteristics of a production technology using production, cost or profit function models. The empirical literature contains a wide variety of models, each underpinned by important

[22] O'Donnell and Coelli (2005) used the M-H algorithm because it is more efficient than the accept-reject algorithm when truncation effects are severe. In such cases the accept-reject algorithm may need to generate extremely large numbers of candidate draws before obtaining just one draw that can be accepted into the sample (eg., Terrell, 1996).

assumptions concerning functional form and the distribution of random errors. The purpose of this chapter is to provide an overview of some common assumptions and their implications for econometric estimation and hypothesis testing. Under fairly weak assumptions it is usually possible and appropriate to estimate models using the method of least squares. Slightly stronger distributional assumptions allow us to estimate the unknown parameters using maximum likelihood or Bayesian techniques. Maximum likelihood estimators are popular because they have desirable large sample properties. Bayesian estimation is becoming increasingly popular, not least because it allows us to obtain exact finite-sample results concerning nonlinear functions of the parameters.

Throughout this chapter we assume that (each equation in) our econometric model contains a symmetric error term representing statistical noise. These models are sometimes known as *average response* models. In the next chapter we introduce one-sided errors that represent inefficiency. Such models are known as *stochastic frontier* models.

9. STOCHASTIC FRONTIER ANALYSIS

9.1 Introduction

In Chapter 5, we consider a measure of the economic efficiency of a firm consisting of two components: technical efficiency, which measures the ability of the firm to obtain the maximum output from given inputs; and allocative efficiency, which measures the ability of the firm to use inputs in optimal proportions given their prices. Computing these efficiency measures involves estimating the unknown production frontier. In Chapters 6 and 7, we obtain a frontier using nonparametric Data Envelopment Analysis (DEA). In this chapter, we consider methods for estimating the frontier parametrically.

Throughout this chapter, we assume, as in Chapter 8, that we have cross-sectional data on I firms. One method for estimating a production frontier using such data is to envelop the data points using an arbitrarily-chosen function. This approach was used by Aigner and Chu (1968) who considered a Cobb-Douglas production frontier of the form:

$$\ln q_i = \mathbf{x}_i'\boldsymbol{\beta} - u_i \qquad i = 1, \ldots, I, \tag{9.1}$$

where q_i represents the output of the i-th firm; $\mathbf{x}_i$ is a $K \times 1$ vector containing the logarithms of inputs; $\boldsymbol{\beta}$ is a vector of unknown parameters; and u_i is a non-negative random variable associated with technical inefficiency. Several techniques can be used to estimate the unknown parameters in this model: Aigner and Chu used linear

programming;[1] Afriat (1972) assumed that the u_is were gamma distributed random variables and used the method of maximum likelihood; while Richmond (1974) used a least squares technique, sometimes known as *modified ordinary least squares (MOLS)*.

The production frontier 9.1 is *deterministic* insofar as q_i is bounded from above by the non-stochastic (i.e., deterministic) quantity $\exp(\mathbf{x}_i'\boldsymbol{\beta})$. A problem with frontiers of this type (and with the DEA frontier) is that no account is taken of measurement errors and other sources of statistical noise – all deviations from the frontier are assumed to be the result of technical inefficiency. Following the discussion in Section 8.3, an obvious solution to the problem is to introduce another random variable representing statistical noise.[2] The resulting frontier is known as a *stochastic* production frontier. This chapter focuses on the estimation of these types of frontiers.

In Section 9.2, we begin with a description of the basic stochastic production frontier model, where (the logarithm of) output is specified as a function of a non-negative random error which represents technical inefficiency, and a symmetric random error which accounts for noise. In Section 9.3, we discuss maximum likelihood estimation of the model and, in Section 9.4, we show how the estimated parameters of the model can be used to predict the technical efficiencies of firms and industries. In Section 9.5, we discuss procedures for testing various hypotheses, including the null hypothesis that all technical inefficiency effects are zero.

Throughout the chapter, we illustrate the various techniques using the Philippine rice data discussed in Appendix 2.

9.2 The Stochastic Production Frontier

Aigner, Lovell and Schmidt (1977) and Meeusen and van den Broeck (1977) independently proposed the stochastic frontier production function model of the form

$$\ln q_i = \mathbf{x}_i'\boldsymbol{\beta} + v_i - u_i \tag{9.2}$$

which is identical to the model 9.1 except we have added a symmetric random error, v_i, to account for statistical noise. Recall from Chapter 8 that statistical noise arises from the inadvertent omission of relevant variables from the vector $\mathbf{x}_i$, as well as

[1] The linear programming problem was to minimise the sum of the $u_i = \ln q_i - \mathbf{x}_i'\boldsymbol{\beta}$ subject to $u_i \geq 0$. Aigner and Chu also suggested the use of quadratic programming.

[2] Another solution is to i) estimate the frontier using all the observations in the sample; ii) delete an arbitrary percentage of the sample firms closest to the frontier; then iii) re-estimate the frontier using the reduced sample. This so-called *probabilistic* frontier approach was suggested by Aigner and Chu (1968) and used by Timmer (1971). The arbitrariness of the selection of a percentage of observations to delete has meant the approach has not been widely adopted in the literature.

from measurement errors and approximation errors associated with the choice of functional form.[3] The model defined by 9.2 is called a *stochastic* frontier production function because the output values are bounded from above by the stochastic (i.e., random) variable $\exp(\mathbf{x}_i'\boldsymbol{\beta} + v_i)$. The random error v_i can be positive or negative and so the stochastic frontier outputs vary about the deterministic part of the model, $\exp(\mathbf{x}_i'\boldsymbol{\beta})$.

These important features of the stochastic frontier model can be illustrated graphically. To do so it is convenient to restrict attention to firms that produce the output q_i using only one input, x_i. In this case, a Cobb-Douglas stochastic frontier model takes the form:

$$\ln q_i = \beta_0 + \beta_1 \ln x_i + v_i - u_i \tag{9.3}$$

or

$$q_i = \exp(\beta_0 + \beta_1 \ln x_i + v_i - u_i) \tag{9.4}$$

or

$$q_i = \underbrace{\exp(\beta_0 + \beta_1 \ln x_i)}_{\substack{\text{deterministic}\\ \text{component}}} \times \underbrace{\exp(v_i)}_{\text{noise}} \times \underbrace{\exp(-u_i)}_{\text{inefficiency}} \tag{9.5}$$

Such a frontier is depicted in Figure 9.1 where we plot the inputs and outputs of two firms, A and B, and where the deterministic component of the frontier model has been drawn to reflect the existence of diminishing returns to scale. Values of the input are measured along the horizontal axis and outputs are measured on the vertical axis. Firm A uses the input level x_A to produce the output q_A, while Firm B uses the input level x_B to produce the output q_B (these observed values are indicated by the points marked with $\times$). If there were no inefficiency effects (i.e., if $u_A = 0$ and $u_B = 0$) then the so-called *frontier* outputs would be

$$q_A^* \equiv \exp(\beta_0 + \beta_1 \ln x_A + v_A) \quad \text{and} \quad q_B^* \equiv \exp(\beta_0 + \beta_1 \ln x_B + v_B)$$

for firms A and B respectively. These frontier values are indicated by the points marked with $\otimes$ in Figure 9.1. It is clear that the *frontier* output for Firm A lies *above* the deterministic part of the production frontier only because the noise effect is *positive* (i.e., $v_A > 0$), while the frontier output for Firm B lies *below* the deterministic part of the frontier because the noise effect is *negative* (i.e., $v_B < 0$). It can also be seen that the *observed* output of Firm A lies *below* the deterministic part of the frontier because the sum of the noise and inefficiency effects is negative (i.e., $v_A - u_A < 0$).

[3] Some authors also use the term 'statistical noise' to refer to the effects of weather, strikes, luck, etc., on the value of the output variable. However, these effects have less to do with our statistical models than with the risky environment in which production takes place. Methods for dealing with risk are discussed in Chapter 10.

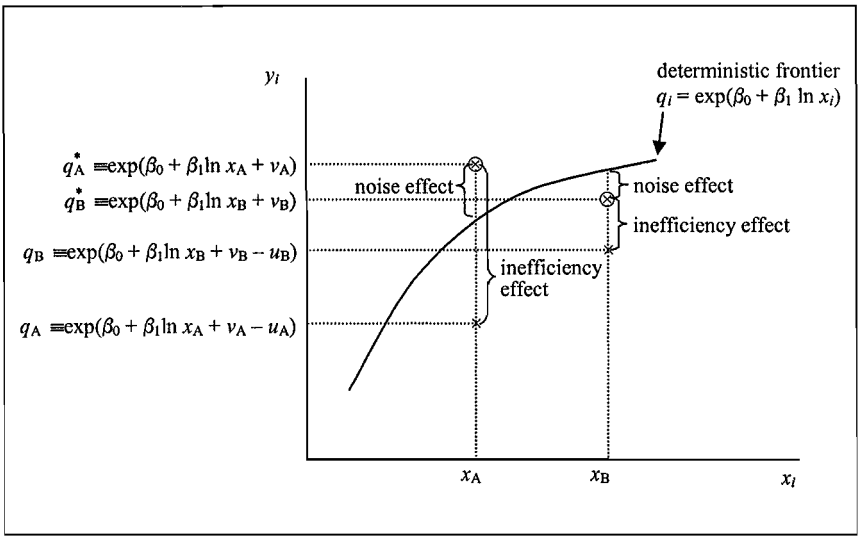

Figure 9.1 The Stochastic Production Frontier

These features of the frontier model 9.3 generalise to the case where firms use several inputs. Specifically, (unobserved) *frontier* outputs tend to be evenly distributed above and below the deterministic part of the frontier. However, *observed* outputs tend to lie below the deterministic part of the frontier. Indeed, they can only lie above the deterministic part of the frontier when the noise effect is positive *and* larger than the inefficiency effect (i.e., $q_i^* > \exp(\mathbf{x}_i'\boldsymbol{\beta})$ iff $\epsilon_i \equiv v_i - u_i > 0$).

Much of stochastic frontier analysis is directed towards the prediction of the inefficiency effects. The most common output-oriented measure of technical efficiency is the ratio of observed output to the corresponding stochastic frontier output:

$$TE_i = \frac{q_i}{\exp(\mathbf{x}_i'\boldsymbol{\beta} + v_i)} = \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta} + v_i - u_i)}{\exp(\mathbf{x}_i'\boldsymbol{\beta} + v_i)} = \exp(-u_i)\,. \tag{9.6}$$

This measure of technical efficiency takes a value between zero and one. It measures the output of the i-th firm relative to the output that could be produced by a fully-efficient firm using the same input vector. Clearly the first step in predicting[4] the technical efficiency, TE_i, is to estimate the parameters of the stochastic production frontier model 9.2.

[4] Because TE_i is a random variable, not a parameter, we use the term "predict" instead of "estimate".

9.3 Estimating the Parameters

Most of the estimation and hypothesis testing procedures discussed in Chapter 8 generalise to the case of stochastic frontiers. Of course, estimation is slightly more complicated due to the fact that the right-hand side of model 9.2 includes two random terms – a symmetric error, v_i, and a non-negative random variable, u_i. As we might expect, our estimation methods are underpinned by assumptions concerning these two random variables.

It is common to assume that each v_i is distributed independently of each u_i and that both errors are uncorrelated with the explanatory variables in $\mathbf{x}_i$. In addition,

$$E(v_i) = 0, \qquad \text{(zero mean)} \qquad (9.7)$$

$$E(v_i^2) = \sigma_v^2, \qquad \text{(homoskedastic)} \qquad (9.8)$$

$$E(v_i v_j) = 0 \text{ for all } i \neq j, \qquad \text{(uncorrelated)} \qquad (9.9)$$

$$E(u_i^2) = \text{constant}, \qquad \text{(homoskedastic)} \qquad (9.10)$$

and $$E(u_i u_j) = 0 \text{ for all } i \neq j. \qquad \text{(uncorrelated)} \qquad (9.11)$$

Thus, the noise component v_i is assumed to have properties that are identical to those of the noise component in the classical linear regression model, discussed in Chapter 8. The inefficiency component has similar properties except it has a non-zero mean (because $u_i \geq 0$).

Under these assumptions, we can obtain consistent estimators of the *slope* coefficients using ordinary least squares (OLS). However, the OLS estimator of the *intercept* coefficient is biased downwards. Among other things, this implies that we cannot use the OLS estimates to compute measures of technical efficiency. One solution to this problem is to correct for the bias in the intercept term using a variant of a method suggested by Winston (1957) – the resulting estimator is often known as the *corrected ordinary least squares (COLS)* estimator.[5] An arguably better solution is to make some distributional assumptions concerning the two error terms and estimate the model using the method of maximum likelihood (ML). Because the ML estimators have many desirable large sample (i.e., asymptotic) properties, they are often preferred to other estimators such as COLS.[6]

9.3.1 The Half-Normal Model

Aigner, Lovell and Schmidt (1977) obtained ML estimates under the assumptions

[5] Winston suggested the COLS estimator in the context of the deterministic frontier 9.1.

[6] The ML estimator is asymptotically more efficient than the COLS estimator. However, the finite-sample properties of the two estimators are unknown. Coelli (1995) provides Monte Carlo evidence suggesting that the ML estimator significantly outperforms the COLS estimator when the contribution of the technical inefficiency effects to the total variance of output is relatively large.

$$v_i \sim iidN(0, \sigma_v^2) \tag{9.12}$$

and $$u_i \sim iidN^+(0, \sigma_u^2). \tag{9.13}$$

Assumption 9.12 says the v_i s are independently and identically distributed normal random variables with zero means and variances σ_v^2. Assumption 9.13 says the u_is are independently and identically distributed half-normal random variables with scale parameter σ_u^2. That is, the probability density function (pdf) of each u_i is a truncated version of a normal random variable having zero mean and variance σ_u^2. To illustrate, the pdfs of three half-normal variables are depicted in Figure 9.2.

Aigner, Lovell and Schmidt (1977) parameterised the log-likelihood function for this so-called half-normal model in terms of $\sigma^2 = \sigma_v^2 + \sigma_u^2$ and $\lambda^2 = \sigma_u^2/\sigma_v^2 \geq 0$. If $\lambda = 0$ there are no technical inefficiency effects and all deviations from the frontier are due to noise.[7] Using this parameterisation, the log-likelihood function is

$$\ln L(\mathbf{y} \mid \boldsymbol{\beta}, \sigma, \lambda) = -\frac{I}{2}\ln\left(\frac{\pi\sigma^2}{2}\right) + \sum_{i=1}^{I} \ln \Phi\left(-\frac{\varepsilon_i \lambda}{\sigma}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{I} \varepsilon_i^2 \tag{9.14}$$

where $\mathbf{y}$ is a vector of log-outputs; $\varepsilon_i \equiv v_i - u_i = \ln q_i - \mathbf{x}_i'\boldsymbol{\beta}$ is a composite error term; and $\Phi(x)$ is the cumulative distribution function (cdf) of the standard normal random variable evaluated at x.

As we know from Chapter 8, maximising a log-likelihood function usually involves taking first-derivatives with respect to the unknown parameters and setting them to zero. Unfortunately, in the case of 9.14 these first-order conditions are highly nonlinear and cannot be solved analytically for $\boldsymbol{\beta}$, σ and λ Thus, we must maximise the likelihood function 9.14 using an iterative optimisation procedure. This involves selecting starting values for the unknown parameters and systematically updating them until the values that maximise the log-likelihood function are found.[8] Details concerning iterative optimisation procedures can be found in Judge *et al.* (1985).

To illustrate ML estimation of the half-normal stochastic frontier model, Table 9.1 presents annotated SHAZAM output from the estimation of a translog production frontier. The model is given by equations 9.2, 9.12 and 9.13 with

[7] Note that λ^2 is not the ratio of the variance of the technical inefficiency effects to the variance of the noise component, since σ_u^2 is the variance of the *untruncated* random variable whose pdf we truncated at zero to obtain the pdf of u_i.

[8] Battese and Corra (1977) found it more convenient to parameterise the log-likelihood in terms of σ^2 and $\gamma = \sigma_u^2/\sigma^2$. The γ-parameter lies between zero and one – if $\gamma = 0$ then all deviations from the frontier are due to noise, while $\gamma = 1$ means all deviations are due to technical inefficiency. This property is convenient for iterative optimisation routines because we can select a starting value by conducting a preliminary search over the unit interval. Coelli (1995) also finds that this parameterisation has computational advantages in COLS estimation.

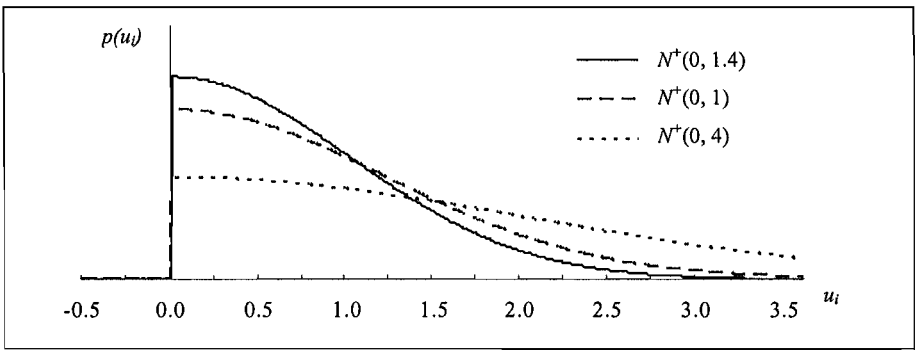

Figure 9.2 Half-Normal Distributions

$$\mathbf{x}_i = \begin{bmatrix} 1 \\ t_i \\ \ln x_{1i} \\ \ln x_{2i} \\ \ln x_{3i} \\ 0.5(\ln x_{1i})^2 \\ \ln x_{1i} \ln x_{2i} \\ \ln x_{1i} \ln x_{3i} \\ 0.5(\ln x_{2i})^2 \\ \ln x_{2i} \ln x_{3i} \\ 0.5(\ln x_{3i})^2 \end{bmatrix}, \tag{9.15}$$

and $$\boldsymbol{\beta} = (\beta_0 \quad \theta \quad \beta_1 \quad \beta_2 \quad \beta_3 \quad \beta_{11} \quad \beta_{12} \quad \beta_{13} \quad \beta_{22} \quad \beta_{23} \quad \beta_{33})', \tag{9.16}$$

where t_i is a time trend included to account for technological change (see Section 8.2.2). Thus, the model is identical to the production function model estimated in Chapter 8 except it includes a time trend and accounts for technical inefficiency. The model is estimated using the same Philippine rice data set used in the empirical examples in Chapter 8. Recall that the variables in this data set have been scaled to have unit means so the first-order coefficients of the translog function can be interpreted as elasticities of output with respect to inputs (evaluated at the variable means). The elasticity estimates reported in Table 9.1 are similar to those reported in Table 8.2 – differences between the two sets of estimates are due to our inclusion of a time trend and, perhaps more importantly, the one-sided random variable representing inefficiency effects.

Table 9.1 Estimating a Half-Normal Frontier Using SHAZAM

```
|_nl 1/ ncoef = 13 logden coef=a
...NOTE..SAMPLE RANGE SET TO:      1,    344
|_eq -0.5*log($pi/2)-0.5*log(sig**2)&
|    +log(ncdf(-(lq-b0-theta*t-b1*lx1-b2*lx2-b3*lx3-b11*lx11-b12*lx12-b13*lx13 &
|    -b22*lx22-b23*lx23-b33*lx33)*lam/sig)) &
|    -((lq-b0-theta*t-b1*lx1-b2*lx2-b3*lx3-b11*lx11-b12*lx12-b13*lx13-b22*lx22 &
|    -b23*lx23-b33*lx33)**2)/(2*sig**2)
|_coef b0 -.04 theta .01 b1 .59 b2 .19 b3 .20 b11 -.44 b12 .68 b13 .06 b22 -.74 &
|      b23 -.18 b33 .02 sig .32 lam .2
  11 VARIABLES IN  1 EQUATIONS WITH  13 COEFFICIENTS
..ALGORITHM USES NUMERIC DERIVATIVES
       344 OBSERVATIONS

< … snip … >

INITIAL STATISTICS :
TIME =       0.021 SEC.    ITER. NO.      0    FUNCT. EVALUATIONS     14
LOG-LIKELIHOOD FUNCTION=    -96.10029
COEFFICIENTS
  0.3200000      -0.4000000E-01  0.1000000E-01  0.5900000       0.1900000
  0.2000000      -0.4400000      0.6800000       0.6000000E-01 -0.7400000
 -0.1800000       0.2000000E-01  0.2000000
GRADIENT
  -8.225026        221.2339       1045.178       -80.26789
  -111.6927        94.05618       175.1684        207.8825
   207.4958        143.3933      -53.29150

INTERMEDIATE STATISTICS :
TIME =       0.312 SEC.    ITER. NO.     15    FUNCT. EVALUATIONS
LOG-LIKELIHOOD FUNCTION=    -83.70183
COEFFICIENTS
  0.4064631       0.2449834       0.1596734E-01  0.5334145
  0.2017508      -0.4490894       0.8735300     -0.1036249       -1.241205
  0.6114771E-01 -0.1401476E-01     1.439884
GRADIENT
   52.39736       -89.14120       -420.9717       5.865800       -7.953317
   14.37761       -15.54588       -15.39514      -37.60257        2.826154
  -25.31484       -35.36377        11.88612
TIME =       0.762 SEC.    ITER. NO.     30    FUNCT. EVALUATIONS    558
LOG-LIKELIHOOD FUNCTION=    -74.40993
COEFFICIENTS
  0.4708611       0.2743642       0.1511094E-01  0.5313797       0.2308970
  0.2032743      -0.4758613       0.6088397       0.6174026E-01 -0.5644706
 -0.1370534      -0.7220351E-02    2.754595
GRADIENT
  0.2977301E-03 -0.1841336E-03 -0.7364420E-03  0.1082024E-04  0.3760791E-04
  0.7039347E-04 -0.3309482E-04 -0.9004729E-04 -0.8957102E-04 -0.6133712E-04
 -0.1249399E-03 -0.7359091E-04  0.5567813E-05

FINAL STATISTICS :

TIME =     0.842 SEC.    ITER. NO.     33    FUNCT. EVALUATIONS
LOG-LIKELIHOOD FUNCTION=    -74.40993
COEFFICIENTS
  0.4708612       0.2743642       0.1511095E-01  0.5313796
  0.2032743      -0.4758611       0.6088395       0.6174011E-01
 -0.1370534      -0.7220231E-02    2.754598
GRADIENT
 -0.1092237E-05  0.2170708E-05  0.1767830E-04 -0.8830360E-05 -0.8381786E-05
 -0.7063150E-05  0.2373536E-05  0.3183649E-05  0.3802588E-05  0.8171241E-06
  0.1961493E-05  0.1840306E-05 -0.1346478E-06
GTRANSPOSE*INVERSE(H)*G  STATISTIC  -  =  0.58130E-13

      COEFFICIENT     ST. ERROR    T-RATIO

SIG       0.47086      0.26677E-01  17.650
B0        0.27436      0.40489E-01   6.7762
THETA     0.15111E-01  0.70054E-02   2.1571
B1        0.53138      0.79547E-01   6.6801
B2        0.23090      0.74768E-01   3.0882
B3        0.20327      0.46905E-01   4.3338
B11      -0.47586      0.20461      -2.3257
B12       0.60884      0.16816       3.6206
B13       0.61740E-01  0.13862       0.44540
B22      -0.56447      0.27054      -2.0864
B23      -0.13705      0.14072      -0.97392
B33      -0.72202E-02  0.93989E-01 -0.76820E-01
LAM        2.7546      0.48687       5.6577
|_end
```

Equation 9.14 with $I = 1$.

Coefficient starting values are arbitrary but should be plausible. When choosing these values we were guided by the OLS results reported in Chapter 8.

After 33 iterations the gradients have collapsed to something close to zero – we have reached a maximum.

The annual percentage change in output due to technological change is estimated to be 1.51% (see Section 8.2.2).

The estimated elasticity of output with respect to labor is 0.23 (when evaluated at the variable means).

It can be seen from Table 9.1 that estimating a frontier model using SHAZAM involves specifying the likelihood function[9]. In practice it is often easier to use purpose-built software packages such as FRONTIER and LIMDEP. For example, the FRONTIER instruction and data files used for estimating the half-normal model are presented in Tables 9.2 and 9.3. The instruction file should be self-explanatory (see the comments on the right-hand side of the file). The data file contains 344 observations, but only four observations are presented in Table 9.3. Each record contains measurements on 13 variables: firm number; year; log-output; and the last 10 variables in the vector $\mathbf{x}_i$ defined by 9.15 (i.e., all variables except the constant term, which is automatically included by FRONTIER). The frontier output file is presented in Table 9.4.

The output generated by LIMDEP is presented in Table 9.5. Lines beginning with the symbol "-->" and ending with "$" are LIMDEP command lines. Notice from Tables 9.3 and 9.4 that the FRONTIER and LIMDEP programs predict firm-specific technical efficiency and inefficiency effects – we will ignore these values here and return to them in Section 9.4. In that Section we see that FRONTIER and LIMDEP are fast and easy to use but they have one important shortcoming – they do not produce prediction intervals for firm-specific or industry technical efficiencies.

Table 9.2 The FRONTIER Instruction File, CHAP9_2.INS

```
1               1=ERROR COMPONENTS MODEL, 2=TE EFFECTS MODEL
chap9.txt       DATA FILE NAME
chap9_2.out     OUTPUT FILE NAME
1               1=PRODUCTION FUNCTION, 2=COST FUNCTION
y               LOGGED DEPENDENT VARIABLE (Y/N)
344             NUMBER OF CROSS-SECTIONS
1               NUMBER OF TIME PERIODS
344             NUMBER OF OBSERVATIONS IN TOTAL
10              NUMBER OF REGRESSOR VARIABLES (Xs)
n               MU (Y/N) [OR DELTA0 (Y/N) IF USING TE EFFECTS MODEL]
n               ETA (Y/N) [OR NUMBER OF TE EFFECTS REGRESSORS (Zs)]
n               STARTING VALUES (Y/N)
```

Table 9.3 The FRONTIER Data File, CHAP9.TXT

```
 1.000000       1.000000       0.1850809      1.000000       0.1538426      0.3961101
 0.9214233E-01  0.1183377E-01  0.6093860E-01  0.1417541E-01  0.7845160E-01  0.3649850E-01
 0.4245104E-02
 2.000000       1.000000       0.4590094      1.000000       0.5725529      0.5296415
 0.4723926      0.1639084      0.3032478      0.2704698      0.1402600      0.2501987
 0.1115774
 3.000000       1.000000       0.4226059      1.000000       0.4613273      0.4505042
 0.2864401      0.1064114      0.2078299      0.1321426      0.1014770      0.1290424
 0.4102396E-01
 4.000000       1.000000      -0.3031307      1.000000      -0.4259759     -0.4657866
-0.7656522      0.9072774E-01  0.1984139      0.3261494      0.1084786      0.3566305
 0.2931116

< … snip … >
```

Firm 3. Year 1. The log-output of firm 3 in year 1

[9] More precisely, it involves specifying the logarithm of the density of a single observation.

Table 9.4 The FRONTIER Output File For The Half-Normal Frontier

```
Output from the program FRONTIER (Version 4.1c)


instruction file = chap9_2.ins
data file =        chap9.txt


 Error Components Frontier (see B&C 1992)
 The model is a production function
 The dependent variable is logged


the ols estimates are :

                  coefficient     standard-error    t-ratio

  beta 0         -0.43313359E-01  0.42960188E-01 -0.10082209E+01
  beta 1          0.12682016E-01  0.77946576E-02  0.16270138E+01
  beta 2          0.58809725E+00  0.85162234E-01  0.69056109E+01
  beta 3          0.19176385E+00  0.80876422E-01  0.23710724E+01
  beta 4          0.19787469E+00  0.51604524E-01  0.38344446E+01
  beta 5         -0.43554684E+00  0.24749127E+00 -0.17598473E+01
  beta 6          0.67864673E+00  0.21659369E+00  0.31332711E+01
  beta 7          0.63920507E-01  0.14561345E+00  0.43897391E+00
  beta 8         -0.74224083E+00  0.30323621E+00 -0.24477315E+01
  beta 9         -0.17828600E+00  0.13861106E+00 -0.12862322E+01
  beta10          0.20367013E-01  0.97907200E-01  0.20802365E+00
  sigma-squared  0.10138434E+00

log likelihood function =  -0.88845085E+02

 < … snip … >

the final mle estimates are :

                  coefficient     standard-error    t-ratio

  beta 0          0.27436347E+00  0.39600416E-01  0.69282978E+01
  beta 1          0.15110945E-01  0.67544802E-02  0.22371736E+01
  beta 2          0.53138167E+00  0.79213877E-01  0.67081892E+01
  beta 3          0.23089543E+00  0.74764329E-01  0.30883101E+01
  beta 4          0.20327381E+00  0.44785423E-01  0.45388387E+01
  beta 5         -0.47586195E+00  0.20221150E+00 -0.23532883E+01
  beta 6          0.60884085E+00  0.16599693E+00  0.36677839E+01
  beta 7          0.61740289E-01  0.13839069E+00  0.44613038E+00
  beta 8         -0.56447322E+00  0.26523510E+00 -0.21281996E+01
  beta 9         -0.13705357E+00  0.14081595E+00 -0.97328160E+00
  beta10         -0.72189747E-02  0.92425705E-01 -0.78105703E-01
  sigma-squared  0.22170997E+00  0.24943636E-01  0.88884383E+01
  gamma          0.88355629E+00  0.36275231E-01  0.24357013E+02
   mu is restricted to be zero
   eta is restricted to be zero

log likelihood function =  -0.74409920E+02


LR test of the one-sided error =   0.28870329E+02
with number of restrictions = 1
 [note that this statistic has a mixed chi-square distribution]

number of iterations =     17

 < … snip … >

technical efficiency estimates :


     firm             eff.-est.

       1            0.77532384E+00
       2            0.72892751E+00
       3            0.77332991E+00

 < … snip … >

     341            0.76900626E+00
     342            0.92610064E+00
     343            0.81931012E+00
     344            0.89042718E+00

 mean efficiency =   0.72941885E+00
```

FRONTIER uses these OLS estimates as starting values for the parameters of the deterministic part of the frontier.

These final estimates are almost identical to those reported by SHAZAM. Different software packages may yield ever-so-slightly different estimates because they may use different iterative optimisation algorithms, starting values and/or convergence criteria.

FRONTIER parameterises the log-likelihood in terms of $\gamma = \sigma_u^2 / \sigma^2$. This estimate (0.88) is high, meaning that much of the variation in the composite error term is due to the inefficiency component.

Table 9.5 Estimating a Half-Normal Frontier Using LIMDEP

```
--> frontier;        Lhs = lq;
                     Rhs = one,t,lx1,lx2,lx3,lx11,lx12,lx13,lx22,lx23,lx33;
                     List; Parameters $
Normal exit from iterations. Exit status=0.

+---------------------------------------------+
| Limited Dependent Variable Model - FRONTIER |
| Maximum Likelihood Estimates                |
| Model estimated: Jan 14, 2005 at 00:46:15PM.|
| Dependent variable                    LQ    |
| Weighting variable                   None   |
| Number of observations                344   |
| Iterations completed                   19   |
| Log likelihood function        -74.40990     |
| Variances: Sigma-squared(v)=        .02582   |
|            Sigma-squared(u)=        .19589   |
|            Sigma(v)        =        .16068   |
|            Sigma(u)        =        .44260   |
| Sigma = Sqr[(s^2(u)+s^2(v)]=        .47086   |
| Stochastic Production Frontier, e=v-u.      |
+---------------------------------------------+
+---------+--------------+----------------+--------+---------+----------+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of X|
+---------+--------------+----------------+--------+---------+----------+
          Primary Index Equation for Model
 Constant     .2743641834   .41063749E-01    6.681   .0000
 T          .1511094895E-01  .66455332E-02    2.274   .0230     4.5000000
 LX1          .5313795262   .83228767E-01    6.385   .0000    -.27185485
 LX2          .2308970687   .85095682E-01    2.713   .0067    -.27723540
 LX3          .2032743669   .51504551E-01    3.947   .0001    -.40784919
 LX11        -.4758606879       .20999250   -2.266   .0234     .35616445
 LX12         .6088388757       .16924939    3.597   .0003     .66448364
 LX13       .6174031369E-01     .14500901     .426   .6703     .78049724
 LX22        -.5644697578       .24230019   -2.330   .0198     .35387530
 LX23        -.1370534724       .13209199   -1.038   .2995     .78880354
 LX33      -.7220275367E-02  .88528934E-01    -.082   .9350     .55289632
          Variance parameters for compound error
 Lambda       2.754597648       .49465604    5.569   .0000
 Sigma        .4708612307   .25808594E-01   18.244   .0000
 (Note: E+nn or E-nn means multiply by 10 to + or -nn power.)

 < … snip … >

Data listing for stochastic frontier model
Observ Data row  Observed Y     Fitted Y       Y - Xb       E[u|e]
     1        1       .1851        .4644       -.2794        .2635
     2        2       .4590        .8211       -.3621        .3264
     3        3       .4226        .7057       -.2830        .2662
     4        4      -.3031       -.2145       -.0886        .1538
     5        5       .2899        .5419       -.2520        .2445
     6        6     -1.2682      -1.2237       -.0445        .1360
     7        7       .1181        .1980       -.0800        .1501
     8        8       .0196        .1204       -.1008        .1592
     9        9       .3412        .5890       -.2478        .2417
    10       10       .3510        .7291       -.3781        .3394
    11       11     -2.0643       -.9855      -1.0788        .9532
    12       12     -1.9614      -1.8685       -.0929        .1556
    13       13      -.5144       -.2895       -.2249        .2267
    14       14       .1302        .5015       -.3713        .3338
    15       15     -1.9614      -1.5125       -.4489        .3986
    16       16     -1.7382       -.9423       -.7959        .7032
    17       17       .5025        .9098       -.4073        .3634
    18       18      1.1699       1.2136       -.0437        .1357
    19       19       .7299        .8470       -.1171        .1667
    20       20       .6267        .7743       -.1476        .1819

 < … snip … >

   335      335     -1.2205       -.3853       -.8351        .7379
   336      336      .4941        .5297       -.0356        .1327
   337      337      .1554        .3367       -.1813        .2003
   338      338      -.7086       -.3858       -.3228        .2956
   339      339      -.5751       -.7717        .1966        .0750
   340      340     -1.4593      -1.1152       -.3441        .3122
   341      341      -.1732        .1178       -.2910        .2719
   342      342      -.2122       -.3851        .1729        .0790
   343      343       .1606        .3530       -.1923        .2067
   344      344       .1671        .1657        .0014        .1201
```

LIMDEP parameterises the log-likelihood in terms of λ.

9.3.2 Other Models

It is not uncommon to replace the half-normality assumption 9.13 with one of the following:

$$u_i \sim iidN^+(\mu, \sigma_u^2) \quad \text{(truncated normal)} \tag{9.17}$$

$$u_i \sim iidG(\lambda, 0) \quad \text{(exponential with mean } \lambda\text{)} \tag{9.18}$$

or $$u_i \sim iidG(\lambda, m). \quad \text{(gamma with mean } \lambda \text{ and degrees of freedom } m\text{)} \tag{9.19}$$

The truncated normal frontier model is due to Stevenson (1980) while the gamma model is due to Greene (1990). The log-likelihood functions for these different models can be found in Kumbhakar and Lovell (2000). Once again, they must be maximised using iterative optimisation routines.

The choice of distributional specification is sometimes a matter of computational convenience – estimation of some frontier models is automated in some software packages but not in others. For example, FRONTIER can be used to estimate half-normal and truncated-normal models, while LIMDEP can also be used to estimate the exponential and gamma models. To illustrate, Table 9.6 presents annotated LIMDEP output from the estimation of our translog production frontier under the assumption that the inefficiency effects are exponentially distributed. A comparison of Tables 9.5 and 9.6 reveals that the estimated elasticities and technological change effects are fairly robust to this change in the distributional assumption.

Theoretical considerations may also influence the choice of the distributional specification. For example, some researchers avoid the half-normal and exponential distributions because they have a mode at zero, implying that most inefficiency effects are in the neighbourhood of zero and the associated measures of technical efficiency would be in the neighbourhood of one. The truncated normal and gamma models allow for a wider range of distributional shapes. To illustrate, Figure 9.3 presents several truncated normal pdfs, two of which can be seen to have non-zero modes. Unfortunately, this sort of flexibility comes at the cost of computational complexity insofar as there are more parameters to estimate. Moreover, if the pdfs for u_i and v_i have similar shapes then it may be difficult to distinguish inefficiency effects from noise.

One final consideration when choosing between models is that different distributional assumptions can give rise to different predictions of technical efficiency. However, when we rank firms on the basis of predicted technical efficiencies, the rankings are often quite robust to distributional choice. In such cases, the principle of parsimony favours the simpler half-normal and exponential models.

Table 9.6 Estimating an Exponential Frontier Using LIMDEP

```
--> frontier;        Lhs = lq;
                     Rhs = one,t,lx1,lx2,lx3,lx11,lx12,lx13,lx22,lx23,lx33;
                     List; Parameters; Model=E $
Normal exit from iterations. Exit status=0.

+-----------------------------------------------+
| Limited Dependent Variable Model - FRONTIER   |
| Maximum Likelihood Estimates                  |
| Model estimated: Jan 18, 2005 at 02:18:41PM.  |
| Dependent variable                    LQ      |
| Weighting variable                   None     |
| Number of observations                344     |
| Iterations completed                   19     |
| Log likelihood function        -71.32557      |
| Exponential frontier model                    |
| Variances: Sigma-squared(v)=        .03514    |
|            Sigma-squared(u)=        .06659    |
| Stochastic Production Frontier, e=v-u.        |
+-----------------------------------------------+
+---------+--------------+----------------+--------+---------+----------+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of X|
+---------+--------------+----------------+--------+---------+----------+
          Primary Index Equation for Model
 Constant     .1828117775   .38792290E-01    4.713   .0000
 T          .1420161846E-01  .63005482E-02    2.254   .0242     4.5000000
 LX1          .5250316451   .76577284E-01    6.856   .0000    -.27185485
 LX2          .2477424253   .79561495E-01    3.114   .0018    -.27723540
 LX3          .2020159208   .45836647E-01    4.407   .0000    -.40784919
 LX11        -.5130865733       .19935997   -2.574   .0101     .35616445
 LX12         .5713640635       .16957890    3.369   .0008     .66448364
 LX13         .1050177508       .12954182     .811   .4175     .78049724
 LX22        -.4596878928       .23972998   -1.918   .0552     .35387530
 LX23        -.1497902887       .12038286   -1.244   .2134     .78880354
 LX33     -.3476481851E-01  .76846954E-01    -.452   .6510     .55289632
          Variance parameters for compound error
 Theta        3.875136294       .41897780    9.249   .0000
 Sigmav       .1874464037   .16441221E-01   11.401   .0000
 (Note: E+nn or E-nn means multiply by 10 to + or -nn power.)

Data listing for stochastic frontier model

 < … snip … >

Observ Data row  Observed Y     Fitted Y      Y - Xb        E[u|e]
     1       1        .1851        .3831      -.1980        .1744
     2       2        .4590        .7360      -.2770        .2137
     3       3        .4226        .6193      -.1967        .1738
     4       4       -.3031       -.3091       .0060        .1081
     5       5        .2899        .4373      -.1473        .1537
     6       6      -1.2682      -1.3165       .0483        .0990
     7       7        .1181        .1006       .0175        .1055
     8       8        .0196        .0317      -.0121        .1124
     9       9        .3412        .5176      -.1764        .1651
    10      10        .3510        .6488      -.2979        .2257
    11      11      -2.0643      -1.0765      -.9878        .8516
    12      12      -1.9614      -1.9338      -.0276        .1163
    13      13       -.5144       -.3838      -.1306        .1476
    14      14        .1302        .4085      -.2783        .2144
    15      15      -1.9614      -1.5732      -.3882        .2853
    16      16      -1.7382      -1.0300      -.7082        .5728
    17      17        .5025        .8496      -.3471        .2566
    18      18       1.1699       1.1484       .0215        .1047
    19      19        .7299        .7698      -.0399        .1195
    20      20        .6267        .7069      -.0802        .1310

 < … snip … >

   335     335      -1.2205       -.4540      -.7664        .6306
   336     336        .4941        .4353       .0588        .0970
   337     337        .1554        .2368      -.0813        .1313
   338     338       -.7086       -.4853      -.2233        .1860
   339     339       -.5751       -.8542       .2791        .0656
   340     340      -1.4593      -1.2009      -.2583        .2036
   341     341       -.1732        .0222      -.1955        .1733
   342     342       -.2122       -.4874       .2752        .0660
   343     343        .1606        .2541      -.0934        .1351
   344     344        .1671        .0633       .1039        .0888
```

The annual percentage change in output due to technological change is estimated to be 1.42% compared to 1.51% for the half-normal model.

The estimated elasticity of output with respect to labor is 0.25 compared to 0.23.

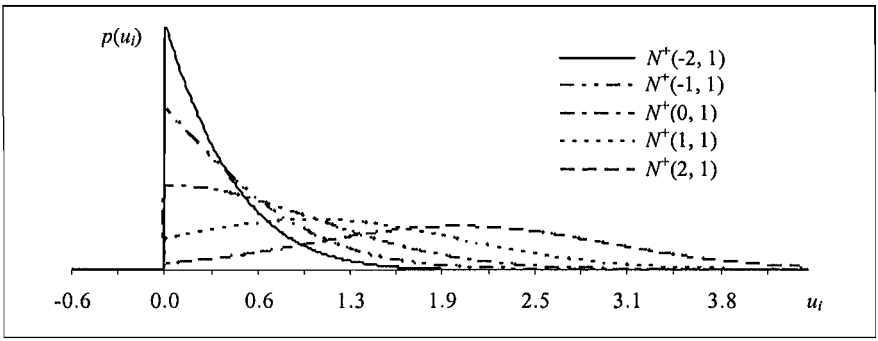

Figure 9.3 Truncated-Normal Distributions

9.4 Predicting Technical Efficiency

Recall, from the end of Section 9.2, that the technical efficiency of the i-th firm is defined by $TE_i = \exp(-u_i)$. This result provides a basis for the prediction of both firm and industry technical efficiency. In this section, we consider these prediction problems in the context of the half-normal stochastic frontier model of Section 9.3.1. Results for other models can be accessed from Kumbhakar and Lovell (2000).

9.4.1 Firm-Specific Efficiency

In order to predict technical efficiency, we clearly need to have some information about the u_is. Once we have collected our data and observed the value of q_i, we can summarise information about u_i in the form of the truncated normal pdf

$$p(u_i|q_i) = \frac{1}{\sqrt{2\pi\sigma_*^2}} \exp\left\{-\frac{1}{2\sigma_*^2}\left(u_i - u_i^*\right)^2\right\} \Big/ \Phi\left(\frac{u_i^*}{\sigma_*}\right) \tag{9.20}$$

where $u_i^* = -(\ln q_i - \mathbf{x}_i'\beta)\sigma_u^2/\sigma^2$ and $\sigma_*^2 = \sigma_v^2\sigma_u^2/\sigma^2$. This conditional pdf gives information about likely and unlikely values of u_i *after* firm i has been selected in our sample and *after* we have observed its output, q_i. Among other things, Jondrow *et al* (1982) use it to derive the following predictor of u_i:

$$\hat{u}_i \equiv E\{u_i|q_i\} = u_i^* + \sigma_* \left[\frac{\phi\left(u_i^*/\sigma_*\right)}{\Phi\left(u_i^*/\sigma_*\right)}\right] \tag{9.21}$$

where $\phi(x)$ is the pdf of the standard normal random variable evaluated at x. Horrace and Schmidt (1995, 1996) also use it to show that a $(1-\alpha)\times100\%$ prediction interval for $u_i \mid q_i$ extends from

$$L_i = u_i^* + \sigma_* \Phi^{-1}\left\{(1-\alpha/2)\Phi\left(u_i^*/\sigma_*\right)\right\} \tag{9.22}$$

to
$$U_i = u_i^* + \sigma_* \Phi^{-1}\left\{(\alpha/2)\Phi\left(u_i^*/\sigma_*\right)\right\}. \tag{9.23}$$

Equations 9.21 to 9.23 are useful because they allow us to draw inferences concerning u_i. However, in most situations, we are more interested in the efficiency of the i-th firm, $TE_i = \exp(-u_i)$. A natural predictor for this quantity is $\exp(-\hat{u}_i)$ where $\hat{u}_i$ is given by equation 9.21. However, Battese and Coelli (1988) have used $p(u_i \mid q_i)$ to derive the alternative predictor

$$T\hat{E}_i \equiv E\left\{\exp(-u_i)\middle|q_i\right\} = \left[\Phi\left(\frac{u_i^*}{\sigma_*} - \sigma_*\right)\Big/\Phi\left(\frac{u_i^*}{\sigma_*}\right)\right]\exp\left\{\frac{\sigma_*^2}{2} - u_i^*\right\}. \tag{9.24}$$

This predictor can be shown to be optimal in the sense that it minimises the mean square prediction error. Irrespective of which of predictor is used, a $(1-\alpha)\times 100\%$ prediction interval is[10]

$$\exp(-U_i) < TE_i < \exp(-L_i) \tag{9.25}$$

where L_i and U_i are given by equations 9.22 and 9.23.

In practice, prediction intervals for firm-specific technical efficiencies can be computed using SHAZAM. To illustrate, Table 9.7 presents both point and interval predictions for the u_is and the TE_is using the half-normal translog production frontier model that is estimated in Section 9.3.1. The predictions for the u_is, which are reported in Table 9.7, are identical to those computed by LIMDEP and reported in Table 9.5. The predictions for the TE_is are the same as those computed by FRONTIER and reported in Table 9.4. Unfortunately, these other software packages do not compute prediction intervals for technical efficiency effects.

9.4.2 Industry Efficiency

Industry efficiency can be viewed as the average of the efficiencies of all the firms in the industry.[11] Thus, a natural predictor of industry efficiency is the average of the predicted efficiencies of the firms in the sample:

$$\overline{TE} \equiv \frac{1}{I}\sum_{i=1}^{I} T\hat{E}_i \tag{9.26}$$

where $T\hat{E}_i$ is computed using equation 9.24.

[10] TE_i is a monotonic transformation of u_i, so lower and upper bounds on u_i translate directly into upper and lower bounds on TE_i.

[11] In some cases a weighted average may be a better measure of industry efficiency if the degree of inefficiency is correlated with firm size in some manner.

Table 9.7 Predicting Firm-Specific Technical Efficiency Using SHAZAM

```
_nl 1/ ncoef = 13 logden coef=a
...NOTE..SAMPLE RANGE SET TO:      1,    344
_eq -0.5*log($pi/2)-0.5*log(sig**2)&
    +log(ncdf(-(lq-b0-theta*t-b1*lx1-b2*lx2-b3*lx3-b11*lx11-b12*lx12-b13*lx13 &
    -b22*lx22-b23*lx23-b33*lx33)*lam/sig)) &
    -((lq-b0-theta*t-b1*lx1-b2*lx2-b3*lx3-b11*lx11-b12*lx12-b13*lx13-b22*lx22 &
    -b23*lx23-b33*lx33)**2)/(2*sig**2)
_coef b0 -.04 theta .01 b1 .59 b2 .19 b3 .20 b11 -.44 b12 .68 b13 .06 b22 -.74 &
      b23 -.18 b33 .02 sig .32 lam .2

< … snip … >

_end
_genl b0 = a:2
_genl theta = a:3
_genl b1 = a:4
_genl b2 = a:5
_genl b3 = a:6
_genl b11 = a:7
_genl b12 = a:8
_genl b13 = a:9
_genl b22 = a:10
_genl b23 = a:11
_genl b33 = a:12
_genl sig = a:1
_genl lam = a:13
_genl sig2 = sig**2
_genl sig2v = sig2/(lam**2+1)
_genl sig2u = sig2v*lam**2
_genl sigu = sqrt(sig2u)
_genr mustar = -(lq-b0-theta*t-b1*lx1-b2*lx2-b3*lx3-b11*lx11-b12*lx12-b13*lx13 &
                 -b22*lx22-b23*lx23-b33*lx33)*sig2u/(sig2)
_genl sigstar = sqrt(sig2u*sig2v/sig2)
_genr musig=mustar/sigstar
_?distrib musig / pdf = pdf1 cdf = cdf1
_genr uhati = mustar + sigstar*pdf1/cdf1
_print obs uhati
      OBS            UHATI
   1.000000       0.2635348
   2.000000       0.3264281
   3.000000       0.2661716
   4.000000       0.1537689
   5.000000       0.2445045
   6.000000       0.1359734
   7.000000       0.1500681

< … snip … >

   340.0000       0.3121585
   341.0000       0.2719096
   342.0000       0.7896287E-01
   343.0000       0.2067270
   344.0000       0.1200562
_genr tei=(ncdf(musig-sigstar)/ncdf(musig))*exp(sigstar**2/2-mustar)
_genr prob1 = 0.975*ncdf(musig)
_genr prob2 = 0.025*ncdf(musig)
_?distrib prob1 / inverse critical = zli
_?distrib prob2 / inverse critical = zui
_genr li = mustar + zli*sigstar
_genr ui = mustar + zui*sigstar
_genr loweri = exp(-ui)
_genr upperi = exp(-li)
_print obs loweri tei upperi
      OBS            LOWERI          TEI            UPPERI
   1.000000       0.5791333       0.7753232       0.9711981
   2.000000       0.5395332       0.7289272       0.9433156
   3.000000       0.5773370       0.7733296       0.9703125
   4.000000       0.6724007       0.8621538       0.9925638
   5.000000       0.5925419       0.7898370       0.9769463
   6.000000       0.6929293       0.8769750       0.9941442
   7.000000       0.6764862       0.8652177       0.9929232
   8.000000       0.6666149       0.8577145       0.9920260

< … snip … >

   341.0000       0.5734757       0.7690060       0.9682988
   342.0000       0.7794616       0.9261005       0.9975220
   343.0000       0.6219603       0.8193098       0.9854132
   344.0000       0.7134347       0.8904269       0.9953080
```

These are exactly same commands used to estimate the frontier in Table 9.1.

For programming purposes it is now convenient to rename the coefficients.

Equation 9.21

LIMDEP also computes these predictions for u_i – see the last column at the bottom of Table 9.3.

Equation 9.24

Equations 9.22 and 9.23

FRONTIER also reports these predictions of firm-specific technical efficiency (but not the prediction interval limits) – see Table 9.2.

Industry efficiency can also be viewed as the expected value of the efficiency of the i-th firm *before* any firms have been selected in the sample. Before we have collected the sample our knowledge of u_i can be summarised in the form of the half-normal pdf[12]

$$p(u_i) = \frac{2}{\sqrt{2\pi\sigma_u^2}} \exp\left\{-\frac{u_i^2}{2\sigma_u^2}\right\}. \tag{9.27}$$

We can use this unconditional pdf to derive results similar to the firm-specific results derived above. Specifically, an optimal estimator of industry efficiency is

$$T\hat{E} \equiv E\{\exp(-u_i)\} = 2\Phi(-\sigma_u)\exp\left\{\frac{\sigma_u^2}{2}\right\}. \tag{9.28}$$

Moreover, a $(1-\alpha)\times 100\%$ prediction interval for industry efficiency is

$$\exp(-U) < TE_i < \exp(-L) \tag{9.29}$$

where $L = z_{.5+\alpha/4}\sigma_u$ and $U = z_{1-\alpha/4}\sigma_u$.

To illustrate, Table 9.8 presents SHAZAM code for evaluating equations 9.28 and 9.29. In this table we use $\alpha = 0.05$, $z_{.5+\alpha/4} = z_{0.5125} = 0.031$ and $z_{1-\alpha/4} = z_{0.9875} = 2.241$. For purposes of comparison, this table also predicts industry efficiency using equation 9.26 (the predictor used by FRONTIER). Prediction intervals are not automatically computed by either LIMDEP or FRONTIER.

Table 9.8 Predicting Industry Technical Efficiency Using SHAZAM

```
< … snip commands from Table 9.4 … >

_genl te = 2*ncdf(-sigu)*exp(sig2u/2)
_genl u = 2.241*sigu
_genl l = 0.031*sigu
_genl lower = exp(-u)
_genl upper = exp(-l)
_print lower te upper
   LOWER
 0.3708850
   TE
 0.7257728
   UPPER
 0.9863731

_stat tei
NAME        N    MEAN       ST. DEV     VARIANCE     MINIMUM     MAXIMUM
TEI         344  0.72942    0.14622     0.21381E-01  0.14997     0.95521
```

Equation 9.28

Equation 9.29

This is $\overline{TE}$ from 9.26. FRONTIER also computes this estimate of industry efficiency – see the bottom of Table 9.2

[12] This is just another way of writing the half-normality assumption (9.13).

9.5 Hypothesis Testing

All the testing procedures discussed in Chapter 8 are available for testing hypotheses concerning β. Recall that the likelihood-ratio (*LR*), Wald and Lagrange Multiplier (*LM*) statistics are only asymptotically justified. Hence, strictly speaking, they can only be relied on when the sample size is large. Unfortunately, the *t*- and *F*-tests discussed in Section 8.5 are no longer justified in small samples because the composed error in the stochastic frontier model is not normally distributed – these tests are also only asymptotically justified.

In addition to testing hypotheses concerning β, stochastic frontier researchers are often interested in testing for the absence of inefficiency effects. In the case of the half-normal and exponential models, the null hypothesis is a single restriction involving a single parameter. If the model has been estimated using the method of maximum likelihood, we can test such an hypothesis using a simple *z*-test (because unconstrained ML estimators are asymptotically normally distributed). For example, in the half-normal model, the null and alternative hypotheses are[13] $H_0 : \sigma_u^2 = 0$ and $H_1 : \sigma_u^2 > 0$ or, if we are using the λ-parameterisation of Aigner, Lovell and Schmidt (1977), $H_0 : \lambda = 0$ and $H_1 : \lambda > 0$. The test statistic is

$$z = \frac{\tilde{\lambda}}{se(\tilde{\lambda})} \sim N(0,1) \tag{9.28}$$

where $\tilde{\lambda}$ is the ML estimator of λ and $se(\tilde{\lambda})$ is the estimator for its standard error. Using the half-normal results reported in Table 9.1, the test statistic is z = 2.755/0.487 = 5.66. This exceeds the critical value $z_{0.95}$ = 1.645 so we reject the null hypothesis that there are no inefficiency effects (at the 5% level of significance).

Coelli (1995) provides Monte Carlo evidence that the *z*-test described above has poor size properties in small samples (i.e., it tends to incorrectly reject the null hypothesis more often than it should). In addition, numerical maximisation of the likelihood function can yield unreliable estimates of covariance matrices and, consequently, standard errors. For these reasons, stochastic frontier researchers often use alternative testing procedures, including Wald and *LR* tests.

Although it is possible to test for inefficiency effects using Wald, *LM* and *LR* tests, the one-sided nature of the alternative hypothesis implies these tests are difficult to interpret. Moreover, they do not have the asymptotic chi-square distributions discussed in Section 8.5. For example, Coelli (1995) shows that the *LR* test statistic given by 8.46 is asymptotically distributed as a mixture of chi-square distributions. In the case of a half-normal model, this means we should reject $H_0 : \lambda = 0$ at the 100α% level of significance if the *LR* test statistic exceeds the

[13] Note that σ_u^2 is not the variance of u_i but rather the variance of the *untruncated* normal random variable whose pdf we truncated at zero. When σ_u^2 = 0 both the untruncated and truncated normal pdfs are concentrated at zero. This implies all the u_is are zero and all firms are fully efficient.

critical value $\chi^2_{1-2\alpha}(1)$. To illustrate, we can use the half-normal results that are reported in Table 9.1 to compute $LR = -2[-88.8451 + 74.4099] = 28.87$ (this value is also reported in Table 9.1). This test statistic exceeds the 5% critical value $\chi^2_{0.9}(1) = 2.71$ so we again reject the null hypothesis of no inefficiency effects. Thus, both the *z*- and *LR*-tests suggest that an average response function is not an adequate representation of the data.

Test procedures of the type described above can also be used to test hypotheses concerning the parameters of other frontier models. For example, in the case of the truncated-normal model, the null hypothesis $H_0 : \mu = \sigma_u^2 = 0$ should be rejected at the 5% level of significance if the *LR* test statistic exceeds 5.138. This value is taken from Table 1 in Kodde and Palm (1986) and is smaller than the 5% critical value, $\chi^2_{0.95}(2) = 5.99$, which was used by several authors including Battese and Coelli (1988). To illustrate, Table 9.9 presents FRONTIER output from the estimation of a truncated-normal model. From the results reported in this table, we compute

$$LR = -2[-88.8451 + 71.6403] = 34.41.$$

This value, which is also reported in Table 9.9, exceeds 5.138 so we reject the null hypothesis.

Finally, we can also use estimates from the truncated-normal model to test the null hypothesis that the simpler half-normal model is adequate. The relevant null and alternative hypotheses are $H_0 : \mu = 0$ and $H_1 : \mu \neq 0$. Again, if the model has been estimated using the method of maximum likelihood, we can use either the *z*- or the *LR*-test. For example, from Table 9.9 we see that the *z*-test statistic has the value $z = -1.537/0.844 = -1.82$. This value exceeds $z_{0.05} = -1.96$ so we do not reject the null hypothesis that the half-normal model is adequate (at the 5% level of significance). Alternatively, from the maximised log-likelihood values reported in Tables 9.4 and 9.9, we calculate the value of the LR statistic as $LR = -2[-74.4099 + 71.6403] = 5.54$. The 5% critical value is $\chi^2_{0.95}(1) = 3.84$. Thus, the LR test leads us to *reject* the null hypothesis that the half-normal model is adequate (at the 5% level of significance). This example illustrates that different testing procedures can lead to different conclusions in finite samples.

Table 9.9 Estimating a Truncated-Normal Frontier Using FRONTIER

```
Output from the program FRONTIER (Version 4.1c)


instruction file = chap9_6.ins
data file =        chap9.txt


 Error Components Frontier (see B&C 1992)
 The model is a production function
 The dependent variable is logged


the ols estimates are :

                  coefficient     standard-error    t-ratio

  beta 0         -0.43313359E-01  0.42960188E-01 -0.10082209E+01
  beta 1          0.12682016E-01  0.77946576E-02  0.16270138E+01
  beta 2          0.58809725E+00  0.85162234E-01  0.69056109E+01
  beta 3          0.19176385E+00  0.80876422E-01  0.23710724E+01
  beta 4          0.19787469E+00  0.51604524E-01  0.38344446E+01
  beta 5         -0.43554684E+00  0.24749127E+00 -0.17598473E+01
  beta 6          0.67864673E+00  0.21659369E+00  0.31332711E+01
  beta 7          0.63920507E-01  0.14561345E+00  0.43897391E+00
  beta 8         -0.74224083E+00  0.30323621E+00 -0.24477315E+01
  beta 9         -0.17828600E+00  0.13861106E+00 -0.12862322E+01
  beta10          0.20367013E-01  0.97907200E-01  0.20802365E+00
  sigma-squared  0.10138434E+00

log likelihood function =  -0.88845085E+02

 < … snip … >

the final mle estimates are :

                  coefficient     standard-error    t-ratio

  beta 0          0.20926836E+00  0.39826659E-01  0.52544794E+01
  beta 1          0.14607027E-01  0.66543281E-02  0.21951168E+01
  beta 2          0.52379627E+00  0.79592579E-01  0.65809687E+01
  beta 3          0.24407896E+00  0.74114218E-01  0.32932811E+01
  beta 4          0.20317928E+00  0.44046162E-01  0.46128715E+01
  beta 5         -0.49268899E+00  0.20662990E+00 -0.23844031E+01
  beta 6          0.58380919E+00  0.16693115E+00  0.34973052E+01
  beta 7          0.84596738E-01  0.13784494E+00  0.61370942E+00
  beta 8         -0.50692510E+00  0.26285864E+00 -0.19285084E+01
  beta 9         -0.13838428E+00  0.13879463E+00 -0.99704347E+00
  beta10         -0.25034657E-01  0.91397678E-01 -0.27390911E+00
  sigma-squared  0.62209328E+00  0.24008893E+00  0.25910952E+01
  gamma          0.94966009E+00  0.25092168E-01  0.37846874E+02
  mu            -0.15372406E+01  0.84375483E+00 -0.18219044E+01
   eta is restricted to be zero

log likelihood function =  -0.71640323E+02

LR test of the one-sided error =   0.34409523E+02
with number of restrictions = 2
 [note that this statistic has a mixed chi-square distribution]


 < … snip … >

technical efficiency estimates :


     firm            eff.-est.

       1           0.82692419E+00
       2           0.79009900E+00
       3           0.82691267E+00

 < … snip … >

     341           0.82642284E+00
     342           0.93535793E+00
     343           0.86372282E+00
     344           0.91143740E+00


 mean efficiency =   0.77471109E+00
```

Maximised value of the (restricted) log-likelihood function when $\mu = \sigma_u^2 = 0$.

Maximised value of the unrestricted log-likelihood function.

9.6 Conclusions

In Chapters 6 and 7 we explain how to estimate production frontiers using the nonparametric DEA approach. One shortcoming of that approach is that it fails to account for statistical noise (eg., the consequences of inadvertently omitting a relevant variable from the production model). In this chapter, we have seen how to overcome this problem using the parametric stochastic frontier approach.

Throughout this chapter, our discussion has focused on maximum likelihood estimation of the parameters of a production frontier and the prediction of individual technical efficiencies. Unfortunately, the simple production frontier model does not permit the prediction of the technical efficiencies of firms that produce multiple outputs. Moreover, the maximum likelihood method does not allow us to assess the reliability of our inferences in small samples. These are some of the issues to be addressed in the next chapter, together with how the parameters of multiple-output technologies can be estimated using distance and cost functions.

10. ADDITIONAL TOPICS ON STOCHASTIC FRONTIER ANALYSIS

10.1 Introduction

In Chapter 9, we discuss methods for estimating single-output production frontier models using cross-sectional data. In this chapter, we extend these ideas to the estimation of distance functions, cost frontiers and production frontiers using panel data. We also move beyond maximum likelihood estimation and consider the estimation of frontier models in a Bayesian framework.

We begin, in Section 10.2, by discussing how distance functions can be used to estimate the parameters of multiple-output production technologies and predict firm-specific and industry technical efficiencies. In Sections 10.3 and 10.4, we consider the estimation of cost frontiers and the decomposition of cost efficiency into its technical and allocative components. In Section 10.5, we consider the relationship between scale efficiency and the elasticity of scale, and show how a measure of scale efficiency can be estimated using a translog production frontier. In Section 10.6, we assume we have access to panel data (i.e., observations on several firms over time) and consider several assumptions concerning the way technical inefficiency effects vary over time. In Section 10.7, we classify characteristics of the production environment as either deterministic (eg., government regulations) or stochastic (eg., weather) and show how these different types of variables can be used to help explain output shortfalls. In Section 10.8, we consider the estimation of frontier models in a Bayesian framework. Once again, throughout the chapter, we illustrate various techniques using the Philippine rice data discussed in Appendix 2.

10.2 Distance Functions

Distance functions can be used to estimate the characteristics of multiple-output production technologies in cases where we have no price information and/or it is inappropriate to assume firms minimise costs or maximise revenues (eg., when the industry is regulated). Input distance functions tend to be used instead of output distance functions when firms have more control over inputs than outputs, and vice versa.[1] To avoid repetition, this section only considers the estimation of input distance functions. For a treatment of output distance functions, see Coelli and Perelman (1999) and O'Donnell and Coelli (2005).

Assume we have access to cross-sectional data on I firms. An input distance function defined over M outputs and N inputs takes the form

$$d_i^I = d^I(x_{1i}, x_{2i}, \ldots x_{Ni}, q_{1i}, q_{2i}, \ldots q_{Mi}) \tag{10.1}$$

where x_{ni} is the n-th input of firm i; q_{mi} is the m-th output; and $d_i^I \geq 1$ is the maximum amount by which the input vector can be radially contracted without changing the output vector (see Chapter 3).[2] Important properties of the function $d^I(.)$ are that it is non-decreasing, linearly homogeneous and concave in inputs, and non-increasing and quasi-concave in outputs.

The first step in econometric estimation of an input distance function is to choose a functional form for $d^I(.)$. Criteria for selecting functional forms are discussed in Section 8.2. It is convenient to choose a functional form that expresses the log-distance as a linear function of (transformations of) inputs and outputs. For example, if we choose the Cobb-Douglas functional form then the model 10.1 becomes

$$\ln d_i^I = \beta_0 + \sum_{n=1}^{N} \beta_n \ln x_{ni} + \sum_{m=1}^{M} \phi_m \ln q_{mi} + v_i \tag{10.2}$$

where v_i is a random variable introduced to account for errors of approximation and other sources of statistical noise (see Section 8.3). This function is non-decreasing, linearly homogeneous and concave in inputs if $\beta_n \geq 0$ for all n and if

$$\sum_{n=1}^{N} \beta_n = 1. \tag{10.3}$$

---

[1] Whether we regard firms as having more control over inputs or outputs may hinge on our definitions of inputs and outputs. For example, hospital patients could be treated as inputs in some modeling contexts but outputs in others.

[2] The superscript I is used in equation 10.1 to indicate that this distance function is an *input* distance function, and should not be confused with the use of I to denote the number of observations. Throughout the book, the interpretation of I should be clear from the context in which it is used.

It is also quasi-concave in outputs if nonlinear functions of the first- and second-order derivatives of d_i^I with respect to the outputs are non-negative. For the reasons discussed in Section 8.7, we ignore these inequality constraints and focus on the equality constraint 10.3.

Econometric estimation of the model 10.2 subject to the constraint 10.3 would be reasonably straightforward were it not for the fact that the dependent variable is unobserved. However, by substituting 10.3 into 10.2 and re-arranging, we obtain an homogeneity-constrained model that is written in the form

$$\ln x_{Ni} = \beta_0 + \sum_{n=1}^{N-1} \beta_n \ln\left(x_{ni}/x_{Ni}\right) + \sum_{m=1}^{M} \phi_m \ln q_{mi} + v_i - u_i \tag{10.4}$$

where $u_i \equiv \ln d_i^I$ is a non-negative variable associated with technical inefficiency. Thus, our decision to expresses $\ln d_i^I$ as a linear function of inputs and outputs results in a model that is in the form of the stochastic production frontier model discussed in Chapter 9. It follows that we can estimate the parameters of the model using the maximum likelihood technique that is discussed in Section 9.3. In addition, a radial input-oriented measure of technical efficiency is

$$TE_i = \frac{1}{d_i^I} = \exp(-u_i). \tag{10.5}$$

Thus, firm-specific and industry technical efficiency can be predicted using the techniques discussed in Section 9.4.

Unfortunately, estimation of distance functions is not always as straightforward as this discussion makes it appear. One issue that concerns some researchers is the possibility that the explanatory variables may be correlated with the composite error term – see Atkinson, Färe and Primont (1998), Atkinson and Primont (1998) and Coelli (2000).[3] This would violate one of the basic assumptions of the stochastic frontier model and would lead to biased estimators. If so, we should estimate the model in an instrumental variables framework. A second issue has to do with the fact that estimated input distance functions often fail to satisfy the concavity and quasi-concavity properties implied by economic theory. This can lead to perverse conclusions regarding the effects of input and output changes on productivity growth and relative efficiency levels. If so, these regularity conditions can be imposed by estimating the model in a Bayesian framework, as presented in O'Donnell and Coelli (2005) and Atkinson and Dorfman (2005).

[3] Coelli (2000) argues that this is not a problem for Cobb-Douglas and translog specifications.

10.3 Cost Frontiers

When price data are available and it is reasonable to assume firms minimise costs, we can estimate the economic characteristics of the production technology (and predict cost efficiency) using a cost frontier. In the case where we have cross-sectional data, the cost frontier model can be written in the general form

$$c_i \geq c(w_{1i}, w_{2i}, \ldots, w_{Ni}, q_{1i}, q_{2i}, \ldots, q_{Mi}) \tag{10.6}$$

where c_i is the observed cost of firm i; w_{ni} is the n-th input price; q_{mi} is the m-th output; and $c(.)$ is a cost function that is non-decreasing, linearly homogeneous and concave in prices. Recall from Chapter 2 that the cost function gives the minimum cost of producing outputs $q_{1i}, q_{2i}, \ldots, q_{Mi}$ when the firm faces prices $w_{1i}, w_{2i}, \ldots, w_{Ni}$. Equation 10.6 says that observed cost is greater than or equal to this minimum cost.

As usual, the first step in estimating the relationship 10.6 is to specify a functional form for $c(.)$. Two convenient choices are the Cobb-Douglas and translog forms. The Cobb-Douglas cost frontier model is

$$\ln c_i \geq \beta_0 + \sum_{n=1}^{N} \beta_n \ln w_{ni} + \sum_{m=1}^{M} \phi_m \ln q_{mi} + v_i \tag{10.7}$$

where v_i is a symmetric random variable representing errors of approximation and other sources of statistical noise. Equivalently:

$$\ln c_i = \beta_0 + \sum_{n=1}^{N} \beta_n \ln w_{ni} + \sum_{m=1}^{M} \phi_m \ln q_{mi} + v_i + u_i \tag{10.8}$$

where u_i is a non-negative variable representing inefficiency. This function is non-decreasing, linearly homogeneous and concave in inputs if the β_n are non-negative and satisfy the constraint

$$\sum_{n=1}^{N} \beta_n = 1. \tag{10.9}$$

Substituting this constraint into the model 10.8 yields the homogeneity-constrained Cobb-Douglas cost frontier model:

$$\ln\left(c_i / w_{Ni}\right) = \beta_0 + \sum_{n=1}^{N-1} \beta_n \ln\left(w_{ni} / w_{Ni}\right) + \sum_{m=1}^{M} \phi_m \ln q_{mi} + v_i + u_i. \tag{10.10}$$

A translog model is obtained in a similar way (see below). Both models are popular in empirical work and can be written in the compact form

$$\ln\left(c_i / w_{Ni}\right) = \mathbf{x}_i'\boldsymbol{\beta} + v_i + u_i \tag{10.11}$$

or, since the distribution of v_i is symmetric,

$$-\ln\left(c_i / w_{Ni}\right) = -\mathbf{x}_i'\boldsymbol{\beta} + v_i - u_i. \tag{10.12}$$

From a statistical viewpoint, equation 10.12 is in exactly the same form as the stochastic production frontier model discussed in Chapter 9. Thus, we can estimate the unknown parameters of the cost frontier using the techniques discussed in Section 9.3. In addition, a measure of cost efficiency is the ratio of minimum cost to observed cost, which can be easily shown to be

$$CE_i = \exp(-u_i). \tag{10.13}$$

Thus, firm-specific and industry cost efficiency can also be predicted using the formulas in Section 9.4.

To illustrate, Table 10.1 presents annotated SHAZAM output from the estimation of a half-normal translog cost frontier defined over a single output and three inputs. Again, the data are the Philippine rice data used in Chapters 8 and 9. After imposing linear homogeneity in prices, the frontier is expressed by[4]

$$\begin{aligned} y_i^* = {} & \beta_0 + \theta t_i + \beta_2 z_{2i} + \beta_3 z_{3i} + \beta_q \ln q_i + \beta_{12} z_{12i} + \beta_{13} z_{13i} + \beta_{23} z_{23i} \\ & + \beta_{q2} z_{q2i} + \beta_{q3} z_{q3i} + \beta_{qq} \left(\ln q_i\right)^2 + v_i + u_i \end{aligned} \tag{10.14}$$

where $y_i^* \equiv \ln c_i - \ln w_{1i}$, (10.15)

$z_{ni} \equiv \ln w_{ni} - \ln w_{1i}$, (10.16)

$z_{nmi} \equiv \ln w_{ni} \ln w_{mi} - 0.5\left(\ln w_{ni}\right)^2 - 0.5\left(\ln w_{mi}\right)^2$, (10.17)

and $z_{qni} \equiv \ln q_i \left(\ln w_{ni} - \ln w_{1i}\right)$. (10.18)

where t_i is a time trend included to account for technological change. It is also possible to estimate this model using LIMDEP and FRONTIER. However, these programs do not compute prediction intervals for firm-specific or industry cost efficiencies – see the discussion in Section 9.4.

[4] Apart from the inclusion of the time trend and the one-sided error representing inefficiency, this model is identical to the cost function model estimated in Section 8.6.

Table 10.1 Estimating a Translog Cost Frontier Using SHAZAM

```
|_nl 1/ ncoef = 13 logden coef=a conv = .1e-7
 ...NOTE..SAMPLE RANGE SET TO:     1,    344
 |_eq -0.5*log($pi/2)-0.5*log(sig**2)&
 |    +log(ncdf(-(-ystar+b0+theta*t+b2*z2+b3*z3+bq*lq+b12*z12+b13*z13 &
 |                +b23*z23+bq2*zq2+bq3*zq3+bqq*lqq)*lam/sig)) &
 |              -((-ystar+b0+theta*t+b2*z2+b3*z3+bq*lq+b12*z12+b13*z13 &
 |                +b23*z23+bq2*zq2+bq3*zq3+bqq*lqq)**2)/(2*sig**2)
 |_coef b0 9.99 theta .03 b2 .33 b3 .12 bq .86 b12 -.11 b13 .01 &
 |       b23 -.01 bq2 -.01 bq3 -.01 bqq .02 sig .32 lam 1.2

 < … snip … >

 FINAL STATISTICS :

 TIME =    2.323 SEC.   ITER. NO.    96   FUNCT. EVALUATIONS  1823
 LOG-LIKELIHOOD FUNCTION=    97.39416
 COEFFICIENTS
   0.2471996       9.668702       0.2815153E-01  0.2427589      0.5654232
   0.9608462     -0.1867709       0.3537091E-01  0.1860834     -0.6299591E-01
   0.3476268E-01  0.1147454E-01   1.528746
 GRADIENT
  -0.1275691E-04  0.2830102E-04  0.2229209E-04 -0.2986928E-04 -0.5820411E-04
   0.2327274E-03 -0.5953767E-04  0.5080381E-05 -0.3400658E-04  0.1448512E-03
   0.1430323E-03 -0.3053593E-03  0.2231223E-05
 GTRANSPOSE*INVERSE(H)*G  STATISTIC  -  =  0.24326E-10

       COEFFICIENT    ST. ERROR   T-RATIO

 SIG      0.24720     0.18540E-01  13.333
 B0        9.6687     0.29746E-01  325.04
 THETA    0.28152E-01 0.50224E-02  5.6052
 B2       0.24276     0.29247E-01  8.3004
 B3       0.56542     0.33227E-01  17.017
 BQ       0.96085     0.17227E-01  55.776
 B12     -0.18677     0.38019E-01 -4.9125
 B13      0.35371E-01 0.39726E-01 0.89037
 B23      0.18608     0.73632E-01  2.5272
 BQ2     -0.62996E-01 0.31941E-01 -1.9723
 BQ3      0.34763E-01 0.34080E-01  1.0200
 BQQ      0.11475E-01 0.12736E-01 0.90095
 LAM       1.5287     0.35091      4.3565
 |_end
 |_gen1 b0 = a:2
 |_gen1 theta = a:3
 |_gen1 b2 = a:4
 |_gen1 b3 = a:5
 |_gen1 bq = a:6
 |_gen1 b12 = a:7
 |_gen1 b13 = a:8
 |_gen1 b23 = a:9
 |_gen1 bq2 = a:10
 |_gen1 bq3 = a:11
 |_gen1 bqq = a:12
 |_gen1 sig = a:1
 |_gen1 lam = a:13
 |_gen1 sig2 = sig**2
 |_gen1 sig2v = sig2/(lam**2+1)
 |_gen1 sig2u = sig2v*lam**2
 |_gen1 sigu = sqrt(sig2u)
 |_genr mustar = -(-ystar+b0+theta*t+b2*z2+b3*z3+bq*lq+b12*z12+b13*z13 &
 |                +b23*z23+bq2*zq2+bq3*zq3+bqq*lqq)*sig2u/(sig2)
 |_gen1 sigstar = sqrt(sig2u*sig2v/sig2)
 |_genr musig=mustar/sigstar
 |_?distrib musig / pdf = pdf1 cdf = cdf1
 |_genr uhati = mustar + sigstar*pdf1/cdf1
 |_print obs uhati
       OBS            UHATI
    1.000000       0.2069826
    2.000000       0.1818350
    3.000000       0.1735828
    4.000000       0.1865170

 < … snip … >

    341.0000       0.1089792
    342.0000       0.4548067E-01
    343.0000       0.1573035
    344.0000       0.6553093E-01
 |_genr cei=(ncdf(musig-sigstar)/ncdf(musig))*exp(sigstar**2/2-mustar)
 |_genr prob1 = 0.975*ncdf(musig)
 |_genr prob2 = 0.025*ncdf(musig)
 |_?distrib prob1 / inverse critical = zli
```

LIMDEP also reports these predictions for u_i

Table 10.1 continued.

```
_?distrib prob2 / inverse critical = zui
_genr li = mustar + zli*sigstar
_genr ui = mustar + zui*sigstar
_genr loweri = exp(-ui)
_genr upperi = exp(-li)
_print obs loweri cei upperi
      OBS          LOWERI          CEI          UPPERI
   1.000000      0.6567046      0.8173323      0.9749868
   2.000000      0.6766772      0.8377616      0.9830702
   3.000000      0.6836744      0.8445564      0.9851648
   4.000000      0.6728129      0.8339266      0.9817675
   5.000000      0.7190614      0.8758590      0.9919982

< … snip … >

   341.0000      0.7503995      0.8993404      0.9950110
   342.0000      0.8597713      0.9563148      0.9986913
   343.0000      0.6982696      0.8580928      0.9886193
   344.0000      0.8172135      0.9379268      0.9978594
_stat cei
NAME          N   MEAN        ST. DEV      VARIANCE     MINIMUM      MAXIMUM
COEF.OF.VARIATION CONSTANT-DIGITS
CEI          344  0.85560     0.69818E-01 0.48746E-02  0.47952      0.95828      0.81602E-01
_genl ce = 2*ncdf(-sigu)*exp(sig2u/2)
_genl u = 2.241*sigu
_genl l = 0.031*sigu
_genl lower = exp(-u)
_genl upper = exp(-l)
_print lower ce upper
    LOWER
  0.6290157
    CE
  0.8541940
    UPPER
  0.9936075
```

Point and interval estimates of firm-specific cost efficiency.

Point and interval estimates of industry cost efficiency.

10.4 Decomposing Cost Efficiency

When we have data on input quantities or cost-shares, cost efficiency can be decomposed into technical and allocative efficiency components. One approach involves estimating a cost frontier together with a subset of cost-share equations along the lines of the cost and cost-share system discussed in Section 8.6. A problem with this approach has to do with the fact that the cost frontier contains an error term representing the combined effects of technical and allocative inefficiency (because both types of inefficiency lead to increased costs) while the share equations contain error terms that represent allocative inefficiency only (because technical inefficiency involves a radial expansion of the input vector and this leaves cost shares unchanged).[5] Unfortunately, it is difficult to explicitly model the relationships between these different error terms without making the equation system highly nonlinear and extremely difficult to estimate – for details see Kumbhakar and Lovell (2000). Thus, in this section, we focus on a slightly different decomposition method developed by Schmidt and Lovell (1979). Their method involves estimating a production frontier together with a subset of the first-order conditions for cost minimisation.[6]

[5] This problem was first raised by Greene (1980) and is now known as the Greene Problem.

[6] Another method for decomposing cost efficiency is proposed by Kopp and Diewert (1982) and refined by Zieschang (1983). Practical implementation of this method requires estimation of a cost

Schmidt and Lovell (1979) found it convenient to work with a single-output Cobb-Douglas production frontier:[7]

$$\ln q_i = \beta_0 + \sum_{n=1}^{N} \beta_n \ln x_{ni} + v_i - u_i. \tag{10.14}$$

Minimising cost subject to this technology constraint is a matter of writing out the Lagrangean, taking first-order derivatives and setting them to zero. Taking the logarithm of the ratio of the first and n-th of these first-order conditions yields:

$$\ln\left(\frac{w_{1i}x_{1i}}{w_{ni}x_{ni}}\right) = \ln\left(\frac{\beta_1}{\beta_n}\right) + \eta_{ni} \qquad \text{for } n = 2, \ldots, N, \tag{10.15}$$

where η_{ni} is a random error term introduced to represent allocative inefficiency.[8] This error is positive, negative or zero depending on whether the firm over-utilises, under-utilises or correctly utilises input 1 relative to input n. A firm is regarded as being allocatively efficient if and only if $\eta_{ni} = 0$ for all n.

Schmidt and Lovell (1979) suggest estimating the N equations represented by 10.14 and 10.15 by the method of maximum likelihood under the (reasonable) assumptions that the v_is, the u_is and the η_{ni}s are independently and identically distributed as univariate normal, half-normal and multivariate normal random variables, respectively, i.e.,

$$v_i \sim iidN(0, \sigma_v^2), \tag{10.16}$$

$$u_i \sim iidN^+(0, \sigma_u^2), \tag{10.17}$$

and $$\boldsymbol{\eta}_i = (\eta_{2i}, \eta_{3i}, \ldots, \eta_{Ni})' \square iidN(\mathbf{0}, \boldsymbol{\Sigma}). \tag{10.18}$$

Under these assumptions, the log-likelihood function is

$$\ln L = I\ln(2r) - \frac{IN}{2}\ln(2\pi) - \frac{I}{2}\ln(\sigma^2) - \frac{I}{2}\ln|\boldsymbol{\Sigma}| + \sum_{i=1}^{I}\ln\Phi\left(-\frac{\varepsilon_i}{\sigma}\sqrt{\frac{\gamma}{1-\gamma}}\right) - \frac{1}{2}\sum_{i=1}^{I}\left[\boldsymbol{\eta}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}_i + \varepsilon_i^2/\sigma^2\right] \tag{10.19}$$

---

function in a single equation framework, followed by numerical solution of many sets of N–1 nonlinear equations (one set for every data observation).

[7] The procedure can be generalized to the multiple-output case using a Cobb-Douglas distance function.

[8] Observe that inputs appear in equation 10.15 in ratio form. Thus, a radial expansion in the input vector (i.e., an increase in technical inefficiency) will not cause a departure from the first-order conditions. However, a change in the input mix (i.e., allocative inefficiency) will clearly cause a departure from the first-order conditions.

where $\varepsilon_i \equiv v_i - u_i = \ln q_i - \beta_0 - \sum_{n=1}^{N} \beta_n \ln x_{ni}$, (10.20)

$$\eta_{ni} = \ln\left(\frac{w_{1i}x_{1i}}{w_{ni}x_{ni}}\right) - \ln\left(\frac{\beta_1}{\beta_n}\right), \tag{10.21}$$

and $r = \sum_{n=1}^{N} \beta_n$ measures returns to scale.

The log-likelihood function 10.19 must be maximised numerically. We can use the estimated production frontier 10.14 to predict $TE_i = \exp(-u_i)$ using the methods discussed in Section 9.4. Of course, the motivation for our work was not to predict technical efficiency but to decompose cost efficiency into its technical and allocative components. For this we need to derive the cost function.

As we know from Section 2.4.1, the Cobb-Douglas functional form has the unusual, but convenient, property that it is self-dual. Schmidt and Lovell (1979) exploit this property and show that the cost function associated with the system 10.14 and 10.15 takes the form

$$\ln c_i = \alpha + \sum_{n=1}^{N}\left(\frac{\beta_n}{r}\right)\ln w_{ni} + \frac{1}{r}\ln q_i - (v_i/r) + (u_i/r) + (A_i - \ln r) \tag{10.22}$$

where

$$A_i = \frac{1}{r}\sum_{n=2}^{N}\beta_n\eta_{ni} + \ln\left[\beta_1 + \sum_{n=2}^{N}\beta_n \exp(-\eta_{ni})\right] \tag{10.23}$$

and α is a non-linear function of the β_n. The term u_i/r on the right-hand side of 10.22 measures the increase in the log-cost due to technical inefficiency, while the term $A_i - \ln r$ measures the increase due to allocative inefficiency. As usual, a measure of cost efficiency is the ratio of minimum cost to observed cost and this can be easily shown to be

$$CE_i = CTE_i \times CAE_i \tag{10.24}$$

where the component $CTE_i = \exp(-u_i/r)$ is due to technical inefficiency, and the component $CAE_i = \exp(\ln r - A_i)$ is due to allocative inefficiency.[9] We can obtain point predictions for CTE_i and CAE_i by substituting predictions for u_i and the η_{ni} into these expressions.[10] In turn, we can predict u_i using equation 9.21 (i.e., the Jondrow *et al.* (1982) estimator) and the η_{ni} using equation 10.15 (i.e., the residuals from the N–2 estimated first-order conditions).

[9] If the technology exhibits constant returns to scale then $r = 1$ and $CTE_i = TE_i = \exp(-u_i)$ and $CAE_i = AE_i \equiv \exp(-A_i)$. Thus, $CE_i = TE_i \times AE_i$, which is the familiar expression from Chapter 3.

[10] Prediction intervals are more difficult to obtain because of the presence of r.

To illustrate the method, Table 10.2 presents annotated SHAZAM output from estimation of a three-input Cobb-Douglas production frontier and decomposition of cost efficiency into its two components. For simplicity, we estimate the production frontier in a single equation framework, although more efficient estimators could be obtained by following the suggestion of Schmidt and Lovell (1979) and estimating the frontier in a seemingly unrelated regression framework.

10.5 Scale Efficiency

Scale efficiency refers to the amount by which productivity can be increased by moving to the *most productive scale size* (MPSS) (see Chapter 3). Thus, to measure scale efficiency, we must have a measure of productivity, and we must have a method for identifying the MPSS.

In the case of a single-input production function, we can measure productivity using the *average product* (*AP*):

$$\mathrm{AP}(x) = \frac{f(x)}{x}. \tag{10.25}$$

Then the MPSS is simply the point at which AP(x) is maximised. The first-order condition for a maximum can be easily rearranged to show that the MPSS is the point where the elasticity of scale, defined in Section 2.2.2, equals one and the firm experiences local constant returns to scale. Thus, to measure scale efficiency, we can set the elasticity of scale to one and solve for the MPSS, denoted x_{*}. Scale efficiency at any input level x is then measured as:[11]

$$\mathrm{SE}(x) = \frac{\mathrm{AP}(x)}{\mathrm{AP}(x_{*})} = \frac{f(x)/f(x_{*})}{x/x_{*}} \leq 1. \tag{10.26}$$

This procedure generalises to the multiple-input case, although a measure of productivity is a little more difficult to conceptualise. One possibility is to think of the input vector $\mathbf{x}$ as one unit of a composite input, so that $k\mathbf{x}$ represents k units of input. Then a measure of productivity is the *ray average product* (RAP):

$$\mathrm{RAP}(k\mathbf{x}) = \frac{f(k\mathbf{x})}{k}. \tag{10.27}$$

Again, it can be shown that the MPSS is the point where the elasticity of scale is equal to one. Thus, to measure scale efficiency, we can set the elasticity of scale to one and solve for the optimal number of units of the composite input, denoted k_{*}. A measure of scale efficiency at input level $k\mathbf{x}$ is then:

[11] Conceptually, this is the same procedure used to measure scale efficiency in Chapter 6. In that chapter, AP(x_{*}) is the slope of a CRS frontier, while AP(x) is the slope of a ray passing through a point on a VRS frontier. The ratio of these two average products is simply written as the ratio of the CRS and VRS DEA technical efficiency scores.

Table 10.2 Decomposing Cost Efficiency Using SHAZAM

```
|_nl 1/ ncoef = 7 logden coef=a
 ...NOTE..SAMPLE RANGE SET TO:      1,    344
 |_eq -0.5*log($pi/2)-0.5*log(sig**2)&
 |     +log(ncdf(-(lq-b0-theta*t-b1*lx1-b2*lx2-b3*lx3)*lam/sig)) &
 |     -((lq-b0-theta*t-b1*lx1-b2*lx2-b3*lx3)**2)/(2*sig**2)
 |_coef b0 -.04 theta .01 b1 .59 b2 .19 b3 .20 sig .32 lam 1.5

 < … snip … >

  FINAL STATISTICS :

 TIME =    0.313 SEC.   ITER. NO.    29   FUNCT. EVALUATIONS   330
 LOG-LIKELIHOOD FUNCTION=   -83.76705
 COEFFICIENTS
   0.4917688      0.2705290      0.1489014E-01  0.3557499      0.3507364
   0.2565321       2.966767
 GRADIENT
   0.2442491E-08   0.000000     -0.5684342E-07 -0.1204310E-09 -0.7479397E-09
    0.000000     -0.4736952E-10
 GTRANSPOSE*INVERSE(H)*G  STATISTIC  -  =  0.92267E-19

       COEFFICIENT    ST. ERROR   T-RATIO

 SIG       0.49177     0.26350E-01  18.663
 B0        0.27053     0.31432E-01  8.6067
 THETA     0.14890E-01 0.52634E-02  2.8290
 B1        0.35575     0.60343E-01  5.8954
 B2        0.35074     0.63176E-01  5.5517
 B3        0.25653     0.35773E-01  7.1711
 LAM        2.9668     0.48976      6.0576
 |_end
 |_gen1 b0 = a:2
 |_gen1 theta = a:3
 |_gen1 b1 = a:4
 |_gen1 b2 = a:5
 |_gen1 b3 = a:6
 |_gen1 sig = a:1
 |_gen1 lam = a:7
 |_gen1 sig2 = sig**2
 |_gen1 sig2v = sig2/(lam**2+1)
 |_gen1 sig2u = sig2v*lam**2
 |_gen1 sigu = sqrt(sig2u)
 |_genr mustar = -(lq-b0-theta*t-b1*lx1-b2*lx2-b3*lx3)*sig2u/sig2
 |_gen1 sigstar = sqrt(sig2u*sig2v/sig2)
 |_genr musig=mustar/sigstar
 |_?distrib musig / pdf = pdf1 cdf = cdf1
 |_genr uhati = mustar + sigstar*pdf1/cdf1
 |_genr tei=(ncdf(musig-sigstar)/ncdf(musig))*exp(sigstar**2/2-mustar)
 |_genr h2 = log(s1/s2)-log(b1/b2)
 |_genr h3 = log(s1/s3)-log(b1/b3)
 |_gen1 r = b1 + b2 + b3
 |_genr ai = (1/r)*(b2*h2+b3*h3) + log(b1 + b2*exp(-h2) + b3*exp(-h3))
 |_genr ctei = exp(-uhati/r)
 |_genr caei = exp(log(r) - ai)
 |_genr cei = ctei*caei
 |_print obs tei ctei caei cei
      OBS            TEI            CTEI           CAEI           CEI
   1.000000       0.7516839      0.7361697      0.9304093      0.6849391
   2.000000       0.7404126      0.7245169      0.9494642      0.6879029
   3.000000       0.7849000      0.7707321      0.9343775      0.7201547
   4.000000       0.8670765      0.8577512      0.8378215      0.7186424

 < … snip … >

   339.0000       0.9362054      0.9321857      0.7034344      0.6557315
   340.0000       0.6608293      0.6431266      0.8376984      0.5387461
   341.0000       0.7711484      0.7563815      0.9158443      0.6927277
   342.0000       0.9260517      0.9212375      0.7758002      0.7146962
   343.0000       0.8261162      0.8141110      0.8493570      0.6914709
   344.0000       0.8980248      0.8910095      0.8012778      0.7139461
 |_stat tei ctei caei cei
 NAME          N   MEAN        ST. DEV      VARIANCE      MINIMUM       MAXIMUM
COEF.OF.VARIATION CONSTANT-DIGITS
 TEI          344  0.72011     0.15320      0.23472E-01  0.12702       0.95745       0.21275
 CTEI         344  0.70594     0.15662      0.24530E-01  0.11600       0.95502       0.22186
 CAEI         344  0.88854     0.78849E-01  0.62171E-02  0.60541       0.99736       0.88739E-01
 CEI          344  0.62218     0.12736      0.16219E-01  0.84232E-01   0.83355       0.20469
```

Predicting technical efficiency using the methods (and commands) discussed in Section 9.4.

Calculating firm-specific cost efficiency using 10.21 to 10.24.

Estimates of industry efficiency.

$$\mathrm{SE}(k\mathbf{x}) = \frac{\mathrm{RAP}(k\mathbf{x})}{\mathrm{RAP}(k.\mathbf{x})} = \frac{f(k\mathbf{x})/f(k.\mathbf{x})}{k/k_*} \leq 1, \tag{10.28}$$

or, if $k = 1$,

$$\mathrm{SE}(\mathbf{x}) = \frac{\mathrm{RAP}(\mathbf{x})}{\mathrm{RAP}(k.\mathbf{x})} = \frac{f(\mathbf{x})/f(k.\mathbf{x})}{k_*} \leq 1. \tag{10.29}$$

In principle, setting the elasticity of scale to one and solving for the MPSS is straightforward. Unfortunately, closed-form solutions are not available for most flexible functional forms. Ray (1998) obtains a solution for a translog functional form and derives the associated measure of scale efficiency. Specifically, if the production frontier takes the form

$$\ln y_i = \beta_0 + \sum_{n=1}^{N} \beta_n \ln x_{ni} + \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \beta_{nm} \ln x_{ni} \ln x_{mi} + v_i - u_i \tag{10.30}$$

then the scale efficiency measure 10.29 becomes

$$\mathrm{SE}(\mathbf{x}) = \exp\left\{\frac{(1-\varepsilon(\mathbf{x}))^2}{2\beta}\right\} \tag{10.31}$$

where

$$\varepsilon(\mathbf{x}) = \sum_{n=1}^{N}\left(\beta_n + \sum_{m=1}^{N} \beta_{nm} \ln x_{mi}\right) \tag{10.32}$$

is the elasticity of scale evaluated at $\mathbf{x}$ and

$$\beta = \sum_{n=1}^{N}\sum_{m=1}^{N} \beta_{nm}. \tag{10.33}$$

If the production frontier is concave in inputs (see Section 2.2.1) then β will be less than zero and the scale efficiency measure given by 10.31 will be less than or equal to one[12]. Unfortunately, the estimated parameters of translog production frontiers are frequently inconsistent with this concavity property (eg., the estimates reported in Table 9.1). The problem can be overcome using methods discussed in Section 8.7.

[12] Concavity is sufficient but not necessary for the scale elasticity measure to be less than or equal to one. The necessary condition is $\beta < 0$.

Finally, Balk (2001) further generalises these results to the case of multiple-input multiple-output technologies (using distance functions).

10.6 Panel Data Models

The frontier models we have discussed to this point are mainly for the analysis of cross-sectional data. In this section, we extend the discussion to the case where panel data are available. Panel data sets usually contain more observations than cross-sectional data sets and so, for this reason alone, we expect to obtain more efficient estimators of the unknown parameters and more efficient predictors of technical efficiencies. Perhaps more importantly, panel data often allow us to:

- relax some of the strong distributional assumptions that were necessary to disentangle the separate effects of inefficiency and noise;
- obtain consistent predictions of technical efficiencies;[13] and
- investigate changes in technical efficiencies (as well as the underlying production technology) over time.

Panel data versions of the Aigner, Lovell and Schmidt (1977) model can be written in the general form

$$\ln q_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + v_{it} - u_{it} \tag{10.34}$$

which is identical to the model 9.2 except we have added a subscript "t" to represent time. If we assume the v_{it}s and the u_{it}s are independently distributed we can estimate the parameters of this model using the methods described in Chapter 9. Indeed, the empirical models estimated in Chapter 9 are models of this type (because the panel nature of the Philippine data was ignored and each error was regarded as being independent of every other error in the model).

Although it is convenient for estimation purposes, a problem with assuming the u_{it}s are independently distributed is that we fail to reap any of the benefits listed above. Moreover, for many industries the independence assumption is unrealistic – all other things being equal, we expect efficient firms to remain reasonably efficient from period to period, and we hope that inefficient firms improve their efficiency levels over time. For these reasons, we need to impose some structure on the inefficiency effects in equation 10.34. It is common to classify different structures according to whether they are time-invariant or time-varying.

[13] Suppose inefficiency effects vary across firms but not across time. In the cross-sectional case, an extra observation means an extra firm-specific inefficiency effect to estimate, and provides no extra information with which to estimate the inefficiency effects we already have (so the variance of our predictor of technical efficiency is unchanged). In the panel data case, increasing the number of time periods (while holding the number of cross-sections fixed) means there is more information with which to estimate the same number of inefficiency effects (so the variance of our predictor should decrease). For details, see Schmidt and Sickles (1984).

10.6.1 Time-Invariant Inefficiency Models

One of the simplest structures we can impose on the inefficiency effects is

$$u_{it} = u_i \qquad i = 1, \ldots, I;\ t = 1, \ldots, T, \tag{10.35}$$

where u_i is treated as either a fixed parameter or a random variable – these models are known as the *fixed effects model* and *random effects model* respectively.

The fixed effects model can be estimated in a standard regression framework using dummy variables.[14] Unfortunately, the estimated model can only be used to measure efficiency relative to the most efficient firm in the sample[15] so our estimates may be unreliable if the number of firms is small.

The random effects model can be estimated using either least squares or maximum likelihood techniques. The least squares approach involves writing the model in the form of the standard error-components model discussed in the panel data literature, then applying Estimated Generalised Least Squares (EGLS).[16] The maximum likelihood approach involves making stronger distributional assumptions concerning the u_is. For example, Pitt and Lee (1981) assumed a half-normal distribution while Battese and Coelli (1988) considered the more general truncated normal distribution:[17]

$$u_i \sim iidN^+(\mu, \sigma_u^2). \tag{10.36}$$

The likelihood function for this model is a generalisation of the likelihood function for the half-normal stochastic frontier model discussed in Chapter 9, and formulas for firm-specific and industry efficiencies are also generalizations of the formulas presented in that chapter. The hypothesis testing procedures discussed in Chapter 9 are also applicable.

In practice, models with time-invariant inefficiency effects can be conveniently estimated using FRONTIER and LIMDEP. To illustrate, Table 10.3 presents annotated FRONTIER output from the estimation of a truncated-normal frontier. Note that significant differences exist between the first-order coefficient estimates reported in this table and those reported in Table 9.6 where no account is taken of the panel nature of the data.

---

[14] The j-th dummy variable is defined as $D_{jit} = 1$ if $j = i$ and 0 otherwise. If we include all I dummy variables in the model we need to drop the constant term (to avoid perfect collinearity).

[15] The fixed-effects predictor for technical efficiency is $TE_i = \exp(\alpha_i - \max_i\{\alpha_i\})$ where α_i is the coefficient of the i-th dummy variable. Prediction intervals for technical efficiency can be obtained using formulas in Horrace and Schmidt (1995, 1996).

[16] For more details on the standard error-components model, see Judge *et al.* (1985).

[17] Battese and Coelli (1988) derived their results for the case of balanced panels (i.e., when every firm is observed in every time period). Battese, Coelli and Colby (1989) generalised the model for the case of an unbalanced data set.

Table 10.3 Truncated-Normal Frontier With Time-Invariant Inefficiency Effects

```
Output from the program FRONTIER (Version 4.1c)

instruction file = chap10_3.ins
data file =        chap10.txt


 Error Components Frontier (see B&C 1992)
 The model is a production function
 The dependent variable is logged

 < … snip … >

the final mle estimates are :

                  coefficient     standard-error    t-ratio

  beta 0          0.17897671E+00  0.82297750E-01  0.21747461E+01
  beta 1          0.14293214E-01  0.68312152E-02  0.20923384E+01
  beta 2          0.69677410E+00  0.91044981E-01  0.76530754E+01
  beta 3          0.93988404E-01  0.79248098E-01  0.11860020E+01
  beta 4          0.17110033E+00  0.53475722E-01  0.31995889E+01
  beta 5         -0.32839402E+00  0.23501356E+00 -0.13973407E+01
  beta 6          0.52706185E+00  0.20412803E+00  0.25820161E+01
  beta 7          0.78112673E-01  0.13991284E+00  0.55829524E+00
  beta 8         -0.61132584E+00  0.28308656E+00 -0.21595015E+01
  beta 9         -0.19660475E+00  0.12322520E+00 -0.15954914E+01
  beta10          0.51404621E-01  0.95334776E-01  0.53920115E+00
  sigma-squared   0.12224251E+00  0.37086791E-01  0.32961200E+01
  gamma           0.38008024E+00  0.18366945E+00  0.20693710E+01
  mu              0.14324897E+00  0.22534081E+00  0.63569317E+00
   eta is restricted to be zero

log likelihood function =  -0.71327846E+02

LR test of the one-sided error =   0.35034479E+02
with number of restrictions = 2
 [note that this statistic has a mixed chi-square distributi

number of iterations =     20

(maximum number of iterations set at :   100)

number of cross-sections =     43

number of time periods =      8

total number of observations =    344

thus there are:      0  obsns not in the panel

 < … snip … >

technical efficiency estimates :


     firm             eff.-est.

       1            0.71492662E+00
       2            0.90931996E+00
       3            0.66864054E+00
       4            0.86393559E+00
       5            0.78873662E+00
       6            0.82547477E+00

 < … snip … >

      40            0.65569408E+00
      41            0.90505280E+00
      42            0.85212116E+00
      43            0.69221394E+00


 mean efficiency =   0.79961497E+00
```

We estimate output has been increasing at a rate of 1.43% per annum due to technological change. This estimate is significantly different from zero at the 5% level (two-tailed test).

The estimated elasticity of output with respect to area is 0.70 (when evaluated at the variable means), much higher than the estimate of 0.59 reported in Table 9.6.

This means every firm is represented in every time period – the panel is 'balanced'. FRONTIER can also handle unbalanced panels.

10.6.2 Time-Varying Inefficiency Models

Time-invariant inefficiency models are somewhat restrictive – we would expect managers to learn from experience and for their technical efficiency levels to change systematically over time (and we would expect these changes to become more noticeable as T gets larger). Two models that allow for time-varying technical inefficiency take the form:

$$u_{it} = f(t) \cdot u_i \tag{10.37}$$

where $f(.)$ is a function that determines how technical inefficiency varies over time:

Kumbhakar (1990): $$f(t) = \left[1 + \exp(\alpha t + \beta t^2)\right]^{-1} \tag{10.38}$$

Battese and Coelli (1992): $$f(t) = \exp\left[\eta(t - T)\right] \tag{10.39}$$

where α, β and η are unknown parameters to be estimated. The Kumbhakar (1990) function lies in the unit interval and can be non-increasing, non-decreasing, concave or convex depending on the signs and magnitudes of α and β. The Battese and Coelli (1992) function involves only one unknown parameter and, partly as a consequence, is less flexible. It has the properties $f(t) \geq 0$ and $f(T) = 1$ and is either non-increasing or non-decreasing depending on the sign of η. However, it is convex for all values of η. These flexibility properties are illustrated in Figure 10.1 where we plot both functions for different values of α, β and η. Unfortunately, a limitation of both functions is that they do not allow for a change in the rank ordering of firms over time – the firm that is ranked n-th at the first time period is always ranked n-th (i.e., if $u_i < u_j$ then $u_{it} = f(t) \cdot u_i \leq f(t) \cdot u_j = u_{jt}$ for all t).

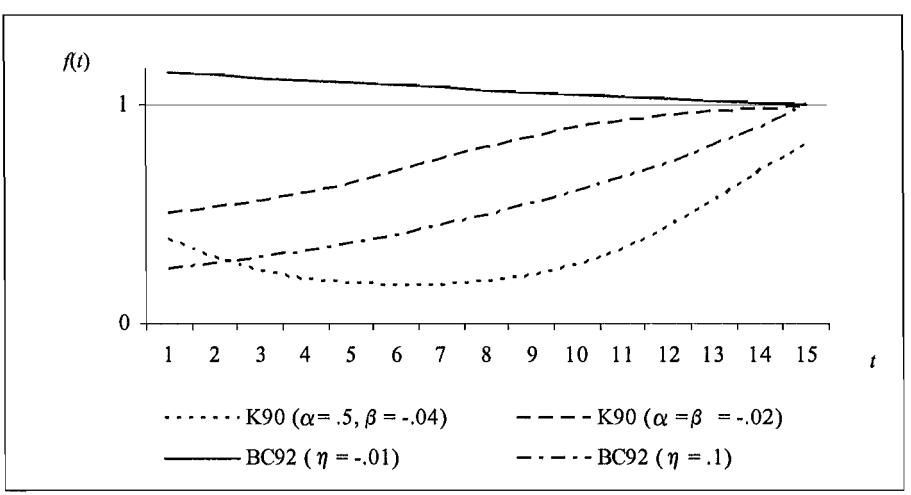

Figure 10.1 Functions for Time-Varying Efficiency Models

Like most frontier models, the models 10.38 and 10.39 can be estimated in a fixed effects framework. However, Kumbhakar (1990) and Battese and Coelli (1992) both propose estimating their models in a random effects framework using the method of maximum likelihood. This often allows us to disentangle the effects of inefficiency and technological change.[18]

The Kumbhakar (1990) and Battese and Coelli (1992) models can both be estimated under the assumption that u_i has a truncated normal distribution:

$$u_i \sim iidN^+(\mu, \sigma_u^2). \tag{10.40}$$

Again, the likelihood function is a generalisation of the likelihood function for the half-normal stochastic frontier model developed in Chapter 9, as are formulas for firm-specific and industry efficiencies. If the model has been estimated by the method of maximum likelihood, hypotheses concerning individual coefficients can be tested using a z test or an LR test. Hypotheses concerning more than one coefficient are usually tested using an LR test. Some null hypotheses of special interest are H_0: $\alpha = \beta = 0$ or H_0: $\eta = 0$ (i.e., time-invariant efficiency effects) and H_0: $\mu = 0$ (i.e., half-normal inefficiency effects at time period T).

To illustrate, Table 10.4 presents annotated FRONTIER output from the estimation of a frontier under assumptions 10.39 and 10.40. Note from this table that we are unable to reject the null hypothesis that the technological change effect is zero using a two-tailed z test at the 5% level of significance. We are also unable to reject H_0: $\eta = 0$ using z or LR tests.[19] This suggests that the model is having difficulty distinguishing between output increases due to technological progress and output increases due to improvements in technical efficiency.

Several more flexible models are discussed in the efficiency literature. For example, Cornwell, Schmidt and Sickles (1990) assume $u_{it} = \beta_{0t} - \theta_{i1} - \theta_{i2}t - \theta_{i3}t^2$, thereby allowing for changes in the rank ordering of firms over time. Lee and Schmidt (1993) write $f(t)$ in equation 10.28 as a linear function of T–1 time dummy variables, implying any temporal pattern in the inefficiency effects is theoretically possible (even implausible patterns with $u_{it} < 0$).[20] Cuesta (2000) specifies a model of the form $u_{it} = \exp\left[\xi_i(t-T)\right]\cdot u_i$. This model generalises the Battese and Coelli (1992) model and allows the temporal pattern of inefficiency effects to vary across firms.

---

[18] In a fixed effects model, all terms involving t end up in the deterministic part of the frontier model, irrespective of what they represent. In a random effects model, some terms involving t end up in the deterministic part of the frontier (these are associated with technological change – see Section 8.2.2) while others feature in the probability density function of u_{it} (these are associated with inefficiency).

[19] The z-test statistic is –0.058 and is reported in Table 10.4. The LR test statistic is computed using results reported in Tables 10.3 and 10.4. Specifically, LR = –2[–71.328 + 70.792] = 1.072. Both statistics are less than their respective critical values at usual levels of significance.

[20] The model involves estimating T–1 dummy variable coefficients so should only be used when T is small.

Table 10.4 Truncated-Normal Frontier With Time-Varying Inefficiency Effects

```
Output from the program FRONTIER (Version 4.1c)

instruction file = chap10_4.ins
data file =        chap10.txt


 Error Components Frontier (see B&C 1992)
 The model is a production function
 The dependent variable is logged

 < … snip … >

the final mle estimates are :

                  coefficient     standard-error    t-ratio

  beta 0          0.23395856E+00  0.94288304E-01  0.24813106E+01
  beta 1         -0.86806565E-03  0.14793311E-01 -0.58679605E-01
  beta 2          0.69441466E+00  0.92839395E-01  0.74797413E+01
  beta 3          0.92355696E-01  0.80496982E-01  0.11473188E+01
  beta 4          0.17203011E+00  0.53139062E-01  0.32373569E+01
  beta 5         -0.27573143E+00  0.23824942E+00 -0.11573226E+01
  beta 6          0.52724921E+00  0.20287301E+00  0.25989126E+01
  beta 7          0.33725992E-01  0.14569943E+00  0.23147649E+00
  beta 8         -0.66602440E+00  0.28654767E+00 -0.23243058E+01
  beta 9         -0.16110214E+00  0.12968777E+00 -0.12422308E+01
  beta10          0.64755268E-01  0.95932923E-01  0.67500569E+00
  sigma-squared   0.10839092E+00  0.30389548E-01  0.35667170E+01
  gamma           0.30469196E+00  0.18903921E+00  0.16117924E+01
  mu              0.68281515E-01  0.22534062E+00  0.30301467E+00
  eta             0.69657401E-01  0.64765677E-01  0.10755296E+01

log likelihood function =  -0.70792099E+02

LR test of the one-sided error =    0.36105972E+02
with number of restrictions = 3
 [note that this statistic has a mixed chi-square distribution]

 < … snip … >

technical efficiency estimates :

efficiency estimates for year      1 :

     firm             eff.-est.

       1            0.68432225E+00
       2            0.89356110E+00
       3            0.62488903E+00

 < … snip … >

      42            0.84902867E+00
      43            0.59751683E+00

 mean eff. in year   1 =  0.76912603E+00


efficiency estimates for year      2 :

     firm             eff.-est.

       1            0.70173930E+00
       2            0.90018362E+00
       3            0.64472171E+00

 < … snip … >
```

The estimated technological change effect is negative but not statistically significant.

This estimate of η is positive, suggesting improvements in technical efficiency over time. However, this effect is not statistically significant.

Firm-specific technical efficiency estimates are increasing over time.

10.7 Accounting for the Production Environment

The ability of a manager to convert inputs into outputs is often influenced by exogenous variables that characterise the environment in which production takes place. When accounting for these variables, it is useful to distinguish between non-stochastic variables that are observable at the time key production decisions are made (eg., degree of government regulation, type of firm ownership, age of the labour force) and unforseen stochastic variables that can be regarded as sources of production risk (eg., weather, pest infestations, events of any type that might lead managers to seek some form of liability insurance).

10.7.1 Non-Stochastic Environmental Variables

Arguably the simplest way to account for non-stochastic environmental variables is to incorporate them directly into the non-stochastic component of the production frontier. In the case of cross-sectional data this leads to a model of the form:

$$\ln q_i = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma} + v_i - u_i \quad , \tag{10.41}$$

where $\mathbf{z}_i$ is a vector of (transformations of) environmental variables and $\boldsymbol{\gamma}$ is a vector of unknown parameters. This model has exactly the same error structure as the conventional stochastic frontier model discussed in Chapter 9. Thus, all the estimators and testing procedures discussed in that chapter are available. For example, if we assume the u_is are half-normally distributed, we can predict the technical efficiency of the i-th firm using equation 9.24:

$$T\hat{E}_i \equiv E\left\{\exp(-u_i)\middle|q_i\right\} = \left[\Phi\left(\frac{u_i^*}{\sigma_*} - \sigma_*\right)\Big/\Phi\left(\frac{u_i^*}{\sigma_*}\right)\right]\exp\left\{\frac{\sigma_*^2}{2} - u_i^*\right\} \tag{9.24}$$

where u_i^* is now a function of both $\mathbf{x}_i$ and $\mathbf{z}_i$. Thus, our predictions of firm-specific technical efficiency now vary with both the traditional inputs *and* the environmental variables. For an empirical example, see Coelli, Perelman and Romano (1999).

Some authors (eg., Pitt and Lee, 1981) explore the relationship between environmental variables and predicted technical efficiencies using a two-stage approach. The first stage involves estimating a conventional frontier model with environmental variables omitted. Firm-specific technical efficiencies are then predicted using formulas such as 9.24. The second stage involves regressing these predicted technical efficiencies on the environmental variables. One problem with this approach is that predicted technical inefficiencies are only a function of environmental variables if the latter are incorporated into the first stage, and doing so makes the second stage unnecessary, because the relationship between the predicted inefficiency effects and the environmental variables is known (it is given

by equations such as 9.24). A related problem is that failure to include environmental variables in the first stage leads to biased estimators of the parameters of the deterministic part of the production frontier, and also to biased predictors of technical efficiency. For more details, see Caudill, Ford and Gropper (1995) and Wang and Schmidt (2002).

A second method for dealing with observable environmental variables is to allow them to directly influence the stochastic component of the production frontier. Kumbhakar, Ghosh and McGuckin (1991) achieve this by assuming

$$\ln q_i = \mathbf{x}_i'\boldsymbol{\beta} + v_i - u_i \quad , \tag{10.42}$$

and $$u_i \sim N^+(\mathbf{z}_i'\boldsymbol{\gamma}, \sigma_u^2). \tag{10.43}$$

Thus, the inefficiency effects in the frontier model have distributions that vary with $\mathbf{z}_i$ so they are no longer identically distributed. The likelihood function is a generalisation of the likelihood function for the conventional model, as are measures of firm-specific and industry efficiency. The model has been generalised to the panel data case by Battese and Coelli (1993, 1995).

Other methods for dealing with observable environmental variables are not always so unambiguously targeted at the deterministic or stochastic components of the frontier. For example, Reifschneider and Stevenson (1991) consider the conventional frontier model 10.42 and assume

$$u_i = g(\mathbf{z}_i) + \varepsilon_i \tag{10.44}$$

where $g(.)$ is a non-negative function (eg., an exponential function) and $\varepsilon_i \sim N^+(0, \sigma_\varepsilon^2)$. This model appears to associate the environmental variables with the inefficiency effects. However, substituting 10.44 into 10.423 yields a model of the form:

$$\ln q_i = \mathbf{x}_i'\boldsymbol{\beta} - g(\mathbf{z}_i) + v_i - \varepsilon_i \tag{10.45}$$

which has exactly the same error structure as a conventional half-normal stochastic frontier. Thus, the Reifschneider and Stevenson model can also be viewed as a model in which the environmental variables influence the deterministic component of the frontier, albeit in a slightly different way to the model 10.41. This suggests an identification problem – we cannot determine whether the environmental variables influence the inefficiency effects or the production technology itself.

10.7.2 Production Risk

A simple way to account for production risk is to append another random variable to the frontier model to represent the combined effects of any variables that are unobserved at the time input decisions are made. If we assume this random variable has a symmetric distribution then it is difficult to distinguish it from the noise v_i. Alternatively, if we assume it has a non-negative distribution it is difficult to distinguish it from the inefficiency effect u_i. This suggests that, for all intents and purposes, we can persist with the conventional stochastic frontier model, although we should recognise that the two error components now measure the effects of noise, inefficiency *and* risk. This is convenient from a statistical viewpoint because we can still use all the estimators and procedures discussed in Chapter 9. However, the conventional frontier model has two undesirable risk properties.

The first of these undesirable properties is that the signs of the marginal products (i.e., derivatives of expected output with respect to inputs) are the same as the signs of the associated marginal risks (i.e., derivatives of the variance of output with respect to inputs). In the context of rainfed rice production, for example, this would mean that higher pesticide use could increase expected output but could not, at the same time, decrease the variance of output. One way of overcoming the problem is to assume the composed error term is heteroskedastic. For example, Battese, Rambaldi and Wan (1997) propose a frontier of the form:[21]

$$q_i = f(\mathbf{x}_i) + g(\mathbf{x}_i)(v_i - u_i) \tag{10.46}$$

where $f(.)$ and $g(.)$ are known functions and v_i represents the combined effects of noise and risk.[22] This model has the property that the marginal risks can be positive or negative depending on the signs of the first-order derivatives of $g(.)$. Battese, Rambaldi and Wan (1997) derive the log-likelihood function under the assumption $u_i \sim iidN^+(\mu, \sigma_u^2)$. As usual, if the model is estimated by the method of maximum likelihood we can test hypotheses concerning the parameters using z and LR tests. The null hypothesis $H_0 : \sigma_u^2 = 0$ is of particular interest – under this hypothesis there are no inefficiency effects and the model collapses to the well-known risk model of Just and Pope (1978).

Chambers and Quiggin (2000) have shown that models such as 10.46 still have a second undesirable risk property, namely they do not permit substitutability between state-contingent outputs. To illustrate what this means, they provide a

[21] This model is a modification of a panel data model developed by Kumbhakar (1993), and has been further generalised by Kumbhakar (2002). Alternative heteroskedastic specifications have been investigated by Caudill and Ford (1993) and Wang (2002).

[22] Battese, Rambaldi and Wan (1997) clearly state that u_i is associated with technical efficiency. By implication, v_i must be associated with sources of risk.

simplified cropping example in which the only input is 'effort' and the only source of production uncertainty is rainfall. In their example, Chambers and Quiggin (2000) assume rainfall is either too little for adequate crop growth (state 1 is drought) or too much for optimal growth (state 2 is flood). Furthermore, a decision must be made about how to allocate effort between the development of irrigation infrastructure and the development of flood-control facilities, and this decision must be made before the amount of rainfall is revealed. It is reasonable to expect that, as more effort is devoted to irrigation infrastructure and less to flood control, output in state 1 (drought) increases and output in state 2 (flood) falls. That is, the firm can trade-off state-contingent outputs by reallocating effort between different activities. Unfortunately, the frontier models we have been discussing up to this point do not allow for this type of substitutability.

One way of allowing for substitution between state-contingent outputs is to estimate a state-contingent stochastic frontier. O'Donnell and Griffiths (2004) show how to do this using a model of the form

$$\ln q_i = \mathbf{x}_i'\boldsymbol{\beta}_j + v_i - u_i \tag{10.47}$$

where $\boldsymbol{\beta}_j$ is a vector of unknown parameters and v_i and u_i represent noise and inefficiency, respectively (not risk). This model is identical to the conventional stochastic frontier model except the coefficient vector $\boldsymbol{\beta}_j$ is permitted to vary across risky states of nature, $j = 1, \ldots, J$ (eg., poor seasonal conditions, average conditions, good conditions). Estimation of such a model is complicated by the fact that states of nature are typically unobserved – although it is possible to measure some characteristics of states of nature (eg., rainfall at germination, humidity at flowering etc.) these data are typically sparse and not in a form that can be used to identify states of nature at the firm level. O'Donnell and Griffiths (2004) overcome the problem by estimating their model in a Bayesian mixtures framework. The estimated model is then used to separately identify output shortfalls due to inefficiency and output shortfalls due to adverse seasonal conditions. In an empirical example using a Philippine rice data set, O'Donnell and Griffiths (2004) found that three-quarters of average estimated output shortfalls experienced by one farmer were due to unfavourable seasonal conditions (i.e., risk) and only one quarter due to inefficiency.

10.8 The Bayesian Approach*

Bayesian estimation is an attractive alternative to maximum likelihood estimation for the reasons discussed in Section 8.8. In this section, we consider Bayesian estimation of the exponential stochastic frontier model mentioned briefly in Section 9.3.2. The model can be written:

* This section contains advanced material that could be optional in an introductory course.

$$\ln q_i = \mathbf{x}_i'\boldsymbol{\beta} + v_i - u_i, \tag{10.48}$$

$$p(\mathbf{v}|h) = (2\pi)^{-I/2} h^{I/2} \exp\left\{-\frac{h}{2}\sum_{i=1}^{I} v_i^2\right\}, \tag{10.49}$$

and $$p(\mathbf{u}|\lambda) = \prod_{i=1}^{I} \lambda^{-1} \exp(-u_i/\lambda) \tag{10.50}$$

where $\mathbf{v} = (v_1, v_2, ..., v_I)'$ and $\mathbf{u} = (u_1, u_2, ..., u_I)'$ are vectors of noise and inefficiency effects; and $h \equiv 1/\sigma_v^2$ is known as the *precision* (for many technical derivations, the precision is easier to work with than the variance). Equations 10.49 and 10.50 are simply alternative ways of writing assumptions 9.12 and 9.18. Our treatment of this model is based on work by van den Broeck, Koop, Osiewalski and Steel (1994), Koop, Steel and Osiewalski (1995) and Koop, Osiewalski and Steel (1997). For a Bayesian treatment of gamma and truncated-normal models see Koop, Steel and Osiewalski (1995).

When estimating frontier models in a Bayesian framework, it is convenient to treat the inefficiency effects as unknown parameters. Then the model 10.48 and the pdf 10.49 imply a likelihood function that is a simple generalisation of equation 8.23 (the likelihood function for the classical linear regression model with normally distributed errors). Specifically:

$$L(\mathbf{y} \mid \boldsymbol{\beta}, h, \mathbf{u}, \lambda) = (2\pi)^{-I/2} h^{I/2} \exp\left\{-\frac{h}{2}\sum_{i=1}^{I} (\ln q_i - \mathbf{x}_i'\boldsymbol{\beta} + u_i)^2\right\} \tag{10.51}$$

where $\mathbf{y} \equiv (\ln q_1, \ln q_2, ..., \ln q_I)'$. In the case of the exponential frontier model it is also convenient to use a prior pdf of the form:

$$p(\boldsymbol{\beta}, h, \mathbf{u}, \lambda) = p(\boldsymbol{\beta}, h) p(\mathbf{u}|\lambda) p(\lambda) \tag{10.52}$$

where $p(\mathbf{u}|\lambda)$ is given by 10.50 and $p(\boldsymbol{\beta}, h)$ and $p(\lambda)$ are chosen by the researcher. Several choices for $p(\boldsymbol{\beta}, h)$ are available, including noninformative and informative priors of the type discussed in Section 8.8.2. For example, if we need to incorporate inequality information into the estimation process, we could use the informative prior

$$p(\boldsymbol{\beta}, h) \propto h \times I(\boldsymbol{\beta}). \tag{10.53}$$

Several choices for $p(\lambda)$ are also available, although this pdf must be proper (an improper prior yields an improper posterior,[23] and this can cause problems for certain aspects of Bayesian inference). One possibility is

$$p(\lambda) = -\ln(\tau^*)\exp\{\lambda^{-1}\ln(\tau^*)\} \tag{10.54}$$

which is the pdf of an exponential random variable with mean $-1/\ln(\tau^*)$. This prior implies a prior median technical efficiency of τ^*. For example, if we think firms in the industry are reasonably efficient we might set $\tau^* = 0.95$.

Combining the likelihood function 10.51 with the prior given by 10.52 to 10.54 yields a joint posterior pdf of non-standard form. However, this joint posterior is of less interest than conditional posterior pdfs that can be derived from it. These conditional pdfs are[24]

$$p(\beta|\mathbf{y},h,\mathbf{u},\lambda) = f_N\left(\beta|\hat{\beta},h^{-1}(\mathbf{X}'\mathbf{X})^{-1}\right)\times I(\beta) \tag{10.55}$$

$$p(h|\mathbf{y},\beta,\mathbf{u},\lambda) = f_G\left(h|I/(\mathbf{y}+\mathbf{u}-\mathbf{X}'\beta)'(\mathbf{y}+\mathbf{u}-\mathbf{X}'\beta),I\right) \tag{10.56}$$

$$p(\lambda^{-1}|\mathbf{y},\beta,h,\mathbf{u}) = f_G\left(\lambda^{-1}|(I+1)/(\mathbf{u}'\mathbf{j}_I-\ln(\tau^*)),2(I+1)\right) \tag{10.57}$$

$$p(u_i|\mathbf{y},\beta,h,\lambda) = f_N\left(u_i|\mathbf{x}_i'\beta-y_i-(h\lambda)^{-1},h^{-1}\right)\times I(u_i) \tag{10.58}$$

where $\mathbf{X} = (\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_I)'$ is an $I \times K$ matrix; $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y}+\mathbf{u})$ is a least squares estimator; and $I(u_i)$ is an indicator function that takes the value 1 if $u_i \geq 0$ and 0 otherwise. It is reasonably straightforward to simulate from these pdfs using the Gibbs sampler discussed in Section 8.9.2. The Gibbs steps involve simulating from the truncated normal pdfs 10.55 and 10.58, and this can be accomplished using accept-reject or Metropolis-Hastings algorithms.

In Table 10.5, we illustrate Bayesian estimation of the model by generating an MCMC sample of size $S = 5{,}500$. To draw from the pdf 10.58, we use a mixture of normal rejection sampling and exponential rejection sampling as suggested by Geweke (1991). The first 500 MCMC observations are discarded as a 'burn-in' and estimates of the parameters of the production technology and measures of firm-specific technical efficiency are obtained by simply averaging over the remaining 5,000 observations. These estimates can be compared with the ML estimates reported in Table 9.4.

---

[23] See Fernandez, Osiewalski and Steel (1997).

[24] The notation $f_N(\mathbf{a}|\mathbf{b},\mathbf{C})$ means $\mathbf{a}$ is a multivariate normal random vector with mean vector $\mathbf{b}$ and covariance matrix $\mathbf{C}$, while $f_G(a|b,c)$ means a is distributed as a gamma random variable with mean b and degrees of freedom parameter c.

Table 10.5 Bayesian Estimation of an Exponential Frontier

```
_smpl 1 1
_set nooutput
_genl iter = 5500
_genl exitcode = 0
_matrix y = lq
_matrix x = constant|t|lx1|lx2|lx3|lx11|lx12|lx13|lx22|lx23|lx33
_matrix b = inv(x'x)*x'y
_matrix sse = (y - x*b)'(y - x*b)
_matrix h = (344-11)/sse
_matrix u = -log(0.875)*constant
_do # = 1,iter
_   matrix varb = (1/h)*inv(x'x)
_   matrix b = inv(x'x)*x'(y+u) + chol(varb)*nor(11,1)
_   matrix sse = (y + u - x*b)'(y + u - x*b)
_   genl mu1 = 2/sse
_   genl urand = uni(1)
_   distrib urand / inverse type = gamma p = mu1 q = 172 critical = h
_   matrix d = u'constant - log(0.875)
_   genl mu2 = 1/d
_   genl urand = uni(1)
_   distrib urand / inverse type = gamma p = mu2 q = 345 critical = linv
_   matrix mu = x*b - y - constant*linv/h
_   do % = 1,344
_      genl check = 0
_      matrix a = -mu(%,1)*sqrt(h)
_      do ! = 1,100
_         genl counter = !
_         genl urand = uni(1)
_         genl z1 = nor(1)
_         genl z2 = a-(1/a)*log(1-uni(1))
_         if1 (z1.ge.a) check = 1
_         if1 ((a.gt.(0.45)).and.(urand.le.exp(-0.5*(z2-a)**2))) check = 2
_         if1 (check.eq.1) newu = (z1 - a)/sqrt(h)
_         if1 (check.eq.2) newu = (z2 - a)/sqrt(h)
_         if1 (counter.eq.100) exitcode = exitcode + 1
_         endif (check.ge.1)
_      endo
_      matrix u(%,1) = newu
_   endo
_   matrix theta = b'|h|linv|u'
_   write (theta.txt) theta / norewind
_endo
_rewind (theta.txt)
_smpl 1 iter
_read (theta.txt) beta0 trend beta1 beta2 beta3 beta11 beta12 beta13 beta22 &
_                 beta23 beta33 prec laminv u1-u344 / eof
_do # = 1,344
_   genr te# = exp(-u#)
_endo
_smpl 501 iter
_set output
_print exitcode
   EXITCODE
  0.000000
_stat beta0 trend beta1 beta2 beta3 beta11 beta12 beta13 beta22 beta23 beta33 &
_     prec laminv
NAME        N   MEAN          ST. DEV       VARIANCE      MINIMUM       MAXIMUM
BETA0       5000  0.17901      0.41321E-01 0.17074E-02  0.43613E-01  0.32105
TREND       5000  0.13680E-01 0.66863E-02 0.44707E-04 -0.90459E-02  0.36696E-01
BETA1       5000  0.53146      0.82040E-01 0.67306E-02  0.22381      0.84435
BETA2       5000  0.24095      0.76553E-01 0.58604E-02 -0.91812E-01  0.51414
BETA3       5000  0.20379      0.45522E-01 0.20723E-02  0.41247E-01  0.36527
BETA11      5000 -0.49499      0.21977     0.48301E-01  -1.3167      0.37802
BETA12      5000  0.57208      0.18414     0.33909E-01 -0.12749       1.3808
BETA13      5000  0.97297E-01 0.14002     0.19606E-01 -0.40034       0.53681
BETA22      5000 -0.45515      0.28557     0.81549E-01  -1.5446      0.69442
BETA23      5000 -0.16443      0.13967     0.19507E-01 -0.62405       0.40941
BETA33      5000 -0.14373E-01 0.92584E-01 0.85717E-02 -0.29353       0.39350
PREC        5000   26.544       5.3566       28.693       10.251       51.304
LAMINV      5000   3.9958      0.47290      0.22363       2.8127       7.6075
_stat te1-te344
NAME        N   MEAN          ST. DEV       VARIANCE      MINIMUM       MAXIMUM
TE1         5000  0.84677      0.10386     0.10787E-01  0.44849      0.99988
TE2         5000  0.81778      0.11235     0.12623E-01  0.41698      0.99998
TE3         5000  0.84509      0.10377     0.10769E-01  0.45750      0.99995

< … snip … >

TE343       5000  0.87685      0.90222E-01 0.81401E-02  0.48911      0.99997
TE344       5000  0.91615      0.69614E-01 0.48461E-02  0.55642      0.99993
```

Prior median technical efficiency is $\tau^* = 0.875$.

Drawing β using 10.45. There are no inequality constraints so $I(\beta) = 1$.

Equation 10.46.

Equation 10.47.

Equation 10.48.

Differences between the estimates reported in Tables 9.4 and 10.5 reflect sampling error and, in the case of the estimated standard errors, the fact that ML estimates do not account for the uncertainty associated with the estimation of the error variances (so they tend to be smaller than the Bayesian estimates).

10.9 Conclusions

In this chapter, we look at two approaches to estimating the parameters of multitple-output technologies. We see that distance functions can be used when no price information is available and/or it is inappropriate to assume that firms minimise costs. Cost frontiers can be used if input prices and output quantities are available and firms are cost minimisers. Two other possible methods for estimating multiple-output technologies are not discussed. First, we can use profit frontiers when input and output prices are available and it is reasonable to assume firms maximise profits. Methods for estimating profit frontiers are similar to those available for estimating cost frontiers – see Berger and Mester (1997). Second, if we are willing to aggregate multiple outputs into a single output measure, we can estimate the technology in a conventional single-output framework. Outputs can be aggregated into a single measure using the index number methods discussed in Chapters 4 and 5.

The decision to estimate a distance function, cost frontier, profit frontier or single-output production frontier is just one of the many decisions facing researchers who want to estimate efficiency using a parametric approach – researchers must also make choices concerning functional forms, error distributions, estimation methods and software. The need to make so many choices is often seen as a disadvantage of the parametric approach. In this context we have two simple pieces of advice. First, always make decisions on a case-by-case basis. Second, whenever it is possible, explore alternative models and estimation methods and (formally or informally) assess the adequacy and robustness of the results obtained.

11. THE CALCULATION AND DECOMPOSITION OF PRODUCTIVITY CHANGE USING FRONTIER METHODS

11.1 Introduction

The principle aim of this chapter is to illustrate how one can, with access to suitable panel data, use the frontier estimation methods discussed in earlier chapters to obtain estimates of TFP change, and decompose these measures into various components, such as technical change and technical efficiency change.

As discussed in Chapter 3, the Malmquist TFP index can be used to measure TFP change for a particular firm between two periods, *s* and *t*.[1] The Malmquist TFP index was defined in that chapter as a ratio of distance function measures. Then in Chapter 4 we explained how one could calculate a Malmquist TFP index using price and quantity data for a single firm (in periods *s* and *t*), if one was willing to make certain assumptions regarding the production technology and the economic behaviour of the firm. These assumptions were required because the limited amount of data available (i.e. on only one firm in the two time periods) prevented us from obtaining an estimate of the production technology in the two time periods.

In this chapter we consider the case where we have access to better data. In particular, we assume we have data on a sample of firms in period *s* and in period t that are sufficient for us to be able to obtain an estimate of the production

[1] It can also be used to calculate TFP change at the industry level or the country level, or alternatively the TFP difference between two firms at one point in time.

technology in these two periods. In this case we can calculate the required distances directly. Hence we can relax a number of the assumptions that were made in constructing the TFP index numbers in Chapter 4. In particular, we need no longer assume that all firms are operating on the surface of the production technology. That is, we no longer need to assume that all firms are technically efficient.[2]

As a result of this we no longer have the situation where the ratio of the distance functions provides a measure of TFP change that is identical to technical change (i.e. frontier shift), as was the case in Chapter 4. Thus, when panel data are available, we can obtain a measure of TFP change that has two components, a technical change component and a technical efficiency change component.

Färe, Grosskopf, Norris and Zhang (1994) take the Malmquist index of TFP growth, defined in Caves, Christensen and Diewert (1982b), and describe how one can decompose the Malmquist TFP change measures into various components, including technical change and efficiency change. They also show how these measures could be calculated using distances measured relative to DEA frontiers.

Although their paper has been very influential, Färe et al (1994) were not the first to specify a TFP change measure that contained both technical change and technical efficiency change components. Nishimizu and Page (1982) estimated translog production frontiers using the Aigner and Chu (1968) linear programming methods and proposed a measure of TFP growth that was the sum of an efficiency change component and a technical change component. However, it should be noted that they did not derive their TFP indices directly using ratios of distances but instead via derivative concepts, similar to those used by Kumbhakar and Lovell (2000) and others. The survey papers by Grosskopf (1993) and Färe, Grosskopf and Roos (1998) provide a good discussion of these two approaches, plus others.

The remainder of this chapter is organised as follows. In Section 11.2, we provide a brief description of the Malmquist TFP index that was first introduced in Chapter 3, and also discuss some additional conceptual issues that are of importance when calculating these indices using panel data. In Section 11.3, we describe how one can calculate these indices using DEA-like methods, similar to those described in Chapters 6 and 7, while in Section 11.4 we describe the calculation of these indices using the SFA methods, described in Chapters 9 and 10. In Section 11.5, we provide a detailed application of some of these methods using the rice industry panel data described in Appendix 2, and in Section 11.6 we make some brief concluding comments.

[2] In addition to the relaxation of this assumption, the assumptions that firms are cost minimisers and/or revenue maximisers are no longer required. This is of particular benefit when analysing public sector and not-for-profit organisations, where these assumptions are unlikely to be valid.

11.2 The Malmquist TFP Index and Panel Data

The description below draws primarily upon the work of Färe *et al.* (1994) and recaps some of the discussion from Chapters 3 and 4. The Malmquist TFP index measures the TFP change between two data points by calculating the ratio of the distances of each data point relative to a common technology. If the period t technology is used as the reference technology, the Malmquist (output-orientated) TFP change index between period s (the base period) and period t is can be written as

$$m_o^t(\mathbf{q}_s,\mathbf{x}_s,\mathbf{q}_t,\mathbf{x}_t)=\frac{d_o^t(\mathbf{q}_t,\mathbf{x}_t)}{d_o^t(\mathbf{q}_s,\mathbf{x}_s)}. \tag{11.1}$$

Alternatively, if the period s reference technology is used it is defined as

$$m_o^s(\mathbf{q}_s,\mathbf{x}_s,\mathbf{q}_t,\mathbf{x}_t)=\frac{d_o^s(\mathbf{q}_t,\mathbf{x}_t)}{d_o^s(\mathbf{q}_s,\mathbf{x}_s)}. \tag{11.2}$$

Note that in the above equations the notation $d_o^s(\mathbf{q}_t,\mathbf{x}_t)$ represents the distance from the period t observation to the period s technology, and all other notation is as previously defined. A value of m_o greater than one indicates positive TFP growth from period s to period t while a value less than one indicates a TFP decline.

As noted by Färe, Grosskopf and Roos (1998), these two (period s and period t) indices are only equivalent if the technology is Hicks output neutral. That is, if the output distance functions may be represented as $d_o^t(\mathbf{q}_t,\mathbf{x}_t)=A(t)\,d_o(\mathbf{q}_t,\mathbf{x}_t)$, for all t. To avoid the necessity to either impose this restriction or to arbitrarily choose one of the two technologies, the Malmquist TFP index is often defined as the geometric mean of these two indices, in the spirit of Fisher (1922) and Caves, Christensen and Diewert (1982b). That is,

$$m_o(\mathbf{q}_s,\mathbf{x}_s,\mathbf{q}_t,\mathbf{x}_t)=\left[\frac{d_o^s(\mathbf{q}_t,\mathbf{x}_t)}{d_o^s(\mathbf{q}_s,\mathbf{x}_s)}\times\frac{d_o^t(\mathbf{q}_t,\mathbf{x}_t)}{d_o^t(\mathbf{q}_s,\mathbf{x}_s)}\right]^{1/2}, \tag{11.3}$$

The distance functions in this productivity index can be rearranged to show that it equivalent to the product of a technical efficiency change index and an index of technical change

$$m_o(\mathbf{q}_s,\mathbf{x}_s,\mathbf{q}_t,\mathbf{x}_t)=\frac{d_o^t(\mathbf{q}_t,\mathbf{x}_t)}{d_o^s(\mathbf{q}_s,\mathbf{x}_s)}\left[\frac{d_o^s(\mathbf{q}_t,\mathbf{x}_t)}{d_o^t(\mathbf{q}_t,\mathbf{x}_t)}\times\frac{d_o^s(\mathbf{q}_s,\mathbf{x}_s)}{d_o^t(\mathbf{q}_s,\mathbf{x}_s)}\right]^{1/2}. \tag{11.4}$$

We observe that the ratio outside the square brackets in the above equation measures the change in the output-oriented measure of Farrell technical efficiency between periods s and t. That is, the efficiency change index is equivalent to the ratio of the Farrell technical efficiency in period t to the Farrell technical efficiency in period s. The remaining part of the index in equation 11.4 is a measure of technical change. It is the geometric mean of the shift in technology between the two periods, evaluated at $\mathbf{x}_t$ and also at $\mathbf{x}_s$. Thus the two terms in equation 11.4 are:

$$\text{Efficiency change} = \frac{d_o^t(\mathbf{q}_t, \mathbf{x}_t)}{d_o^s(\mathbf{q}_s, \mathbf{x}_s)} \tag{11.5}$$

and

$$\text{Technical change} = \left[\frac{d_o^s(\mathbf{q}_t, \mathbf{x}_t)}{d_o^t(\mathbf{q}_t, \mathbf{x}_t)} \times \frac{d_o^s(\mathbf{q}_s, \mathbf{x}_s)}{d_o^t(\mathbf{q}_s, \mathbf{x}_s)}\right]^{1/2}. \tag{11.6}$$

A number of additional possible decompositions of these technical efficiency change and technical change components have been proposed by various authors. Some of these options are discussed in the Färe, Grosskopf and Roos (1998) survey paper. Such as the Färe and Grosskopf (1996) method of decomposing the technical change component into input bias, output bias and "magnitude" components, and the Färe *et al.* (1994) suggestion that technical efficiency change be decomposed into scale efficiency and "pure" technical efficiency components (this can only be done when the distance functions in the above equations are estimated relative to a CRS technology).

This Färe *et al.* (1994) decomposition involving scale efficiency has been widely used and (more recently) also widely criticised. The decomposition involves taking the efficiency change measure in equation 11.5 (which we now need to assume involves the ratio of two CRS distance functions) and decomposing it into a *pure efficiency change* component (measured relative to the arguably *true* VRS frontier)

$$\text{Pure efficiency change} = \frac{d_{ov}^t(\mathbf{q}_t, \mathbf{x}_t)}{d_{ov}^s(\mathbf{q}_s, \mathbf{x}_s)}, \tag{11.7}$$

and a *scale efficiency change* component

$$\text{Scale efficiency change} = \left[\frac{d_{ov}^t(\mathbf{q}_t, \mathbf{x}_t)/d_{oc}^t(\mathbf{q}_t, \mathbf{x}_t)}{d_{ov}^t(\mathbf{q}_s, \mathbf{x}_s)/d_{oc}^t(\mathbf{q}_s, \mathbf{x}_s)} \times \frac{d_{ov}^s(\mathbf{q}_t, \mathbf{x}_t)/d_{oc}^s(\mathbf{q}_t, \mathbf{x}_t)}{d_{ov}^s(\mathbf{q}_s, \mathbf{x}_s)/d_{oc}^s(\mathbf{q}_s, \mathbf{x}_s)}\right]^{1/2}. \tag{11.8}$$

The scale efficiency change component in equation 11.8 is actually the geometric mean of two scale efficiency change measures. The first is relative to the period t technology and the second is relative to the period s technology. Note that the extra subscripts, v and c, relate to the VRS and CRS technologies, respectively. Also note that if this extra decomposition is used, the distance functions in equations 11.1 to 11.6 would all need to be relative to a CRS technology, and hence have an extra c-subscript appended.

The above-suggested method of introducing a scale efficiency change component in the Malmquist TFP index decomposition has been the source of considerable debate in recent years. The main point of criticism is essentially that if there is scale efficiency change then this implies that the *true* production technology must be VRS. However, the Färe *et al.* (1994) decomposition reports a technical change measure (see equation 11.6) that reflects the movement in a CRS frontier and not the VRS frontier. Ray and Desli (1997) point out this inconsistency and suggest an alternative decomposition that has technical change measured relative to VRS frontiers and an amended scale change component (that is no longer equivalent to scale efficiency change).

These two alternative decompositions will be approximately equal if the rate of technical change is similar at the observed data point and at the corresponding most productive scale size (MPSS) point, but will differ otherwise. The Ray and Desli (1997) decomposition is arguably a more "internally consistent" decomposition. However, the differences between the two approaches will only be substantive when there are firms within the sample with significantly different scales, and there are scale economies, and there are non-neutral rates of technical change across the different sized firms. Furthermore, the Ray and Desli (1997) method can suffer from computational difficulties when DEA-based distance functions are used because of infeasibilities in some inter-period VRS calculations. Hence, both approaches have their particular advantages and disadvantages. For further discussion of these and other related issues see Ray and Desli (1997) and the survey paper by Balk (2003).

One important point that is closely related to this issue, is that the returns to scale properties of the technology are very important in TFP measurement. Grifell-Tatjé and Lovell (1995) use a simple one-input, one-output example to illustrate that the Malmquist TFP index defined in equation 11.3 may not correctly measure TFP changes when VRS is assumed for the technology. Hence it is important that CRS be imposed upon the technology that is used to estimate distance functions for the calculation of this Malmquist TFP index, or alternatively that an appropriate adjustment factor is included to correct for this omission.[3] Otherwise the resulting measures may not properly reflect the TFP gains or losses resulting from scale effects.

[3] Orea (2002) suggests the inclusion of an additional scale change component in a Malmquist TFP index derived from a translog technology. This method is discussed in the next section.

11.3 Calculation using DEA Frontiers

There are a number of different methods that could be used to estimate a production technology and, hence, measure the distance functions that make up the Malmquist TFP index. To date, the most popular method has been the DEA-like linear programming methods suggested by Färe *et al.* (1994), which are discussed now in this section. The other main alternative approach is the use of stochastic frontier methods, which are described in section 11.4.

Following Färe *et al.* (1994), and given that suitable panel data are available, we can calculate the distance measures in equation 11.3 using DEA-like linear programs. For the i-th firm, we must calculate four distance functions to measure the TFP change between two periods. This requires the solving of four linear programming (LP) problems. As noted above, Färe *et al.* (1994) utilise a constant returns to scale (CRS) technology in their TFP calculations. This ensures that resulting TFP change measures satisfy the fundamental property that if all inputs are multiplied by the (positive) scalar δ and all outputs are multiplied by the (non-negative) scalar α, then the resulting TFP change index will equal α/δ. The required LPs are:[4]

$$
\begin{aligned}
&[d_o^t(\mathbf{q}_t, \mathbf{x}_t)]^{-1} = \max_{\phi,\lambda} \phi, \\
&\text{st} \quad -\phi\mathbf{q}_{it} + \mathbf{Q}_t\boldsymbol{\lambda} \geq \mathbf{0}, \\
&\qquad \mathbf{x}_{it} - \mathbf{X}_t\boldsymbol{\lambda} \geq \mathbf{0}, \\
&\qquad \boldsymbol{\lambda} \geq \mathbf{0},
\end{aligned} \tag{11.9}
$$

$$
\begin{aligned}
&[d_o^s(\mathbf{q}_s, \mathbf{x}_s)]^{-1} = \max_{\phi,\lambda} \phi, \\
&\text{st} \quad -\phi\mathbf{q}_{is} + \mathbf{Q}_s\boldsymbol{\lambda} \geq \mathbf{0}, \\
&\qquad \mathbf{x}_{is} - \mathbf{X}_s\boldsymbol{\lambda} \geq \mathbf{0}, \\
&\qquad \boldsymbol{\lambda} \geq \mathbf{0},
\end{aligned} \tag{11.10}
$$

$$
\begin{aligned}
&[d_o^t(\mathbf{q}_s, \mathbf{x}_t)]^{-1} = \max_{\phi,\lambda} \phi, \\
&\text{st} \quad -\phi\mathbf{q}_{is} + \mathbf{Q}_t\boldsymbol{\lambda} \geq \mathbf{0}, \\
&\qquad \mathbf{x}_{is} - \mathbf{X}_t\boldsymbol{\lambda} \geq \mathbf{0}, \\
&\qquad \boldsymbol{\lambda} \geq \mathbf{0},
\end{aligned} \tag{11.11}
$$

and

$$
\begin{aligned}
&[d_o^s(\mathbf{q}_t, \mathbf{x}_t)]^{-1} = \max_{\phi,\lambda} \phi, \\
&\text{st} \quad -\phi\mathbf{q}_{it} + \mathbf{Q}_s\boldsymbol{\lambda} \geq \mathbf{0}, \\
&\qquad \mathbf{x}_{it} - \mathbf{X}_s\boldsymbol{\lambda} \geq \mathbf{0}, \\
&\qquad \boldsymbol{\lambda} \geq \mathbf{0}.
\end{aligned} \tag{11.12}
$$

[4] All notation follows directly from that used in Chapters 6 and 7. The only differences are that we now have time subscripts, s and t, to represent the two time periods of interest.

Note that in LP's 11.11 and 11.12, where production points are compared with technologies from different time periods, the ϕ parameter need not be greater than or equal to one, as it must be when calculating Farrell output-orientated technical efficiencies. The data point could lie above the feasible production set. This will most likely occur in LP 11.12 where a production point from period t is compared with technology in an earlier period, s. If technical progress has occurred, then a value of $\phi<1$ is possible. Note that it could also possibly occur in LP 11.11 if technical regress has occurred, but this is less likely.

It is important to note that the ϕs and λs are likely to take different values in the above four LPs. Furthermore, the above four LPs must be solved for each firm in the sample. Thus, if there are 20 firms and two time periods, 80 LPs must be solved. Note also that as extra time periods are added, one must solve an extra three LP's for each firm (to construct a chained index). If there are T time periods, then ($3T$-2) LPs must be solved for each firm in the sample. Hence, if there are I firms, then there are $I\times(3T\text{-}2)$ LPs to be solved. For example, with I=20 firms and T=10 time periods, this would involve $20\times(3\times10\text{-}2) = 560$ LPs.

As noted in the previous section, the above approach can be extended by decomposing the technical efficiency change measure into a scale efficiency measure and a "pure" technical efficiency measure (refer to equations 11.7 and 11.8). This requires the solution of two additional LPs (when comparing two production points). These would involve repeating LPs 11.9 and 11.10 with the convexity restriction ($\mathbf{I1}'\lambda=1$) added to each. This provides estimates of distance functions relative to a variable returns to scale (VRS) technology. For the case of I firms and T time periods, this would increase the number of LPs to be performed from $I\times(3T\text{-}2)$ to $I\times(4T\text{-}2)$.

A Simple Numerical Example

Since these DEA-like methods have been used in the vast majority of studies that have used panel data to construct Malmquist TFP indices, we give a simple numerical example using the DEAP computer program.

In this example, we take the data from the simple one-output, one-input example introduced in Chapter 6 (see Table 6.6a) and add an extra year of data. These data are listed in Table 11.1 and are also plotted in Figure 11.2. Also plotted in Figure 11.2 are the CRS and VRS DEA frontiers for the two time periods.

The text file EG4-DTA.TXT (refer to Table 11.2a) contains observations on five firms over a three-year period. These firms produce one output using one input. Data for year 1 are listed in the first five rows, year 2 data are in the second five rows and year 3 data are listed in the final five rows. Note that the year 1 and 2 data are identical to those listed in Table 11.1. Note also that the year 2 data are identical to the year 3 data. This is done to ensure that the example remains quite simple.

Table 11.1 Example Data for Malmquist DEA

firm	year	q	x
1	1	1	2
2	1	2	4
3	1	3	3
4	1	5	5
5	1	5	6
1	2	1	2
2	2	3	4
3	2	4	3
4	2	5	5
5	2	5	5

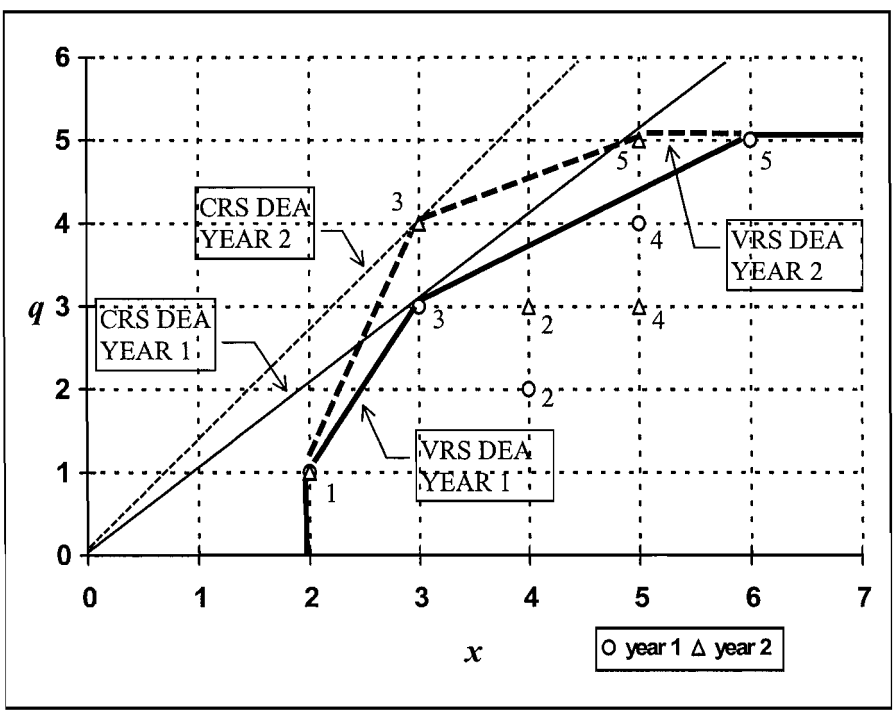

Figure 11.1 Malmquist DEA Example

Table 11.2a Listing of Data File, EG4-DTA.TXT

```
1 2
2 4
3 3
4 5
5 6
1 2
3 4
4 3
3 5
5 5
1 2
3 4
4 3
3 5
5 5
```

The EG4-INS.TXT file is listed in Table 11.2b. The only changes relative to the EG2-INS.TXT instruction file (which was used in our VRS DEA example) listed in Table 6.6b is that:

- the input and output file names are different;
- the number of time periods is now 3;
- a "1" is entered on the third last line to indicate that an output orientation is required; and
- a "2" is entered on the last line to indicate that Malmquist DEA is required.

Note that the VRS/CRS option in the DEAP instruction file has no influence on the Malmquist DEA routine because both are used to calculate the various distances that are used to construct the Malmquist indices.

Table 11.2b Listing of Instruction File, EG4-INS.TXT

```
eg4-dta.txt     DATA FILE NAME
eg4-out.txt     OUTPUT FILE NAME
5               NUMBER OF FIRMS
3               NUMBER OF TIME PERIODS
1               NUMBER OF OUTPUTS
1               NUMBER OF INPUTS
1               0=INPUT AND 1=OUTPUT ORIENTATED
0               0=CRS AND 1=VRS
2               0=DEA(MULTI-STAGE), 1=COST-DEA, 2=MALMQUIST-DEA,
                       3=DEA(1-STAGE), 4=DEA(2-STAGE)
```

The output file, EG4-OUT.TXT, is reproduced in Table 11.2c. The output begins with a listing of the distances needed for the Malmquist calculations. Four distances are calculated for each firm in each year. These are relative to:

- the previous period's CRS DEA frontier;
- the current period's CRS DEA frontier;
- the next period's CRS DEA frontier; and
- the current period's VRS frontier.

Following this the Malmquist indices are presented. All indices are calculated relative to the previous year. Hence the output begins with year 2. Five indices are presented for each firm in each year. These are:

- technical efficiency change (relative to a CRS technology);
- technological change;
- pure technical efficiency change (i.e., relative to a VRS technology);
- scale efficiency change; and
- total factor productivity (TFP) change.

Following this, summary tables of these indices are presented for the different time periods (over all firms) and for the different firms (over all time periods). Note that all indices are equal to one for time period 3. This is because, in the example data set used (see Table 11.2a), the data for year 3 are identical to the year 2 data.

Table 11.2c Listing of Output File, EG4-OUT.TXT

```
Results from DEAP Version 2.1

Instruction file    = eg4-ins.txt
Data file           = eg4-dta.txt

Output orientated Malmquist DEA

 DISTANCES SUMMARY

 year =      1

  firm       crs te rel to tech in yr       vrs
   no.       ************************        te
              t-1          t        t+1

     1      0.000      0.500      0.375      1.000
     2      0.000      0.500      0.375      0.545
     3      0.000      1.000      0.750      1.000
     4      0.000      0.800      0.600      0.923
     5      0.000      0.833      0.625      1.000

 mean       0.000      0.727      0.545      0.894
```

year = 2

firm no.	crs te rel to tech in yr t-1	t	t+1	vrs te
1	0.500	0.375	0.375	1.000
2	0.750	0.563	0.563	0.667
3	1.333	1.000	1.000	1.000
4	0.600	0.450	0.450	0.600
5	1.000	0.750	0.750	1.000
mean	0.837	0.628	0.628	0.853

year = 3

firm no.	crs te rel to tech in yr t-1	t	t+1	vrs te
1	0.375	0.375	0.000	1.000
2	0.563	0.563	0.000	0.667
3	1.000	1.000	0.000	1.000
4	0.450	0.450	0.000	0.600
5	0.750	0.750	0.000	1.000
mean	0.628	0.628	0.000	0.853

[Note that t-1 in year 1 and t+1 in the final year are not defined]

MALMQUIST INDEX SUMMARY

year = 2

firm	effch	techch	pech	sech	tfpch
1	0.750	1.333	1.000	0.750	1.000
2	1.125	1.333	1.222	0.920	1.500
3	1.000	1.333	1.000	1.000	1.333
4	0.562	1.333	0.650	0.865	0.750
5	0.900	1.333	1.000	0.900	1.200
mean	0.844	1.333	0.955	0.883	1.125

year = 3

firm	effch	techch	pech	sech	tfpch
1	1.000	1.000	1.000	1.000	1.000
2	1.000	1.000	1.000	1.000	1.000
3	1.000	1.000	1.000	1.000	1.000
4	1.000	1.000	1.000	1.000	1.000
5	1.000	1.000	1.000	1.000	1.000
mean	1.000	1.000	1.000	1.000	1.000

MALMQUIST INDEX SUMMARY OF ANNUAL MEANS

year	effch	techch	pech	sech	tfpch
2	0.844	1.333	0.955	0.883	1.125
3	1.000	1.000	1.000	1.000	1.000
mean	0.918	1.155	0.977	0.940	1.061

```
MALMQUIST INDEX SUMMARY OF FIRM MEANS

 firm    effch   techch    pech     sech    tfpch

    1    0.866    1.155   1.000    0.866    1.000
    2    1.061    1.155   1.106    0.959    1.225
    3    1.000    1.155   1.000    1.000    1.155
    4    0.750    1.155   0.806    0.930    0.866
    5    0.949    1.155   1.000    0.949    1.095

mean     0.918    1.155   0.977    0.940    1.061

[Note that all Malmquist index averages are geometric means]
```

11.4 Calculation using SFA Frontiers

The distance measures required for the Malmquist TFP index calculations can also be measured relative to a parametric technology. A number of papers have been written in recent years that describe ways in which this can be done. The majority of these can be classified into two groups: those that derive the measures using derivative-based techniques (e.g. see Kumbhakar and Lovell, 2000) and those that seek to use explicit distance measures (e.g. see Fuentes, Grifell-Tatjé and Perelman, 2001). The two approaches tend to provide TFP formulae and decompositions that are quite similar. Hence, in this chapter we confine our attention to the latter approach, mostly to maintain consistency with our Malmquist index concepts.

The methods we present in this section are based upon the translog distance function methods described in Fuentes, Grifell-Tatjé and Perelman (2001) and Orea (2002). However, we focus our attention on the production frontier case, which is a single-output special case of the more general (multi-output) output distance function.

We consider a translog stochastic production frontier defined as follows

$$\begin{aligned}\ln q_{it} = \beta_0 &+ \sum_{n=1}^{N} \beta_n \ln x_{nit} + \frac{1}{2}\sum_{n=1}^{N}\sum_{j=1}^{N} \beta_{nj} \ln x_{nit} \ln x_{nit} \\ &+ \sum_{n=1}^{N} \beta_{tn} t \ln x_{nit} + \beta_t t + \frac{1}{2}\beta_{tt} t^2 + v_{it} - u_{it}, \\ &\qquad i=1,2,\ldots,I,\ t=1,2,\ldots,T, \end{aligned} \tag{11.13}$$

where q_{it} is the output of the i-th firm in the t-th year;

x_{nit} denotes a n-th input variable;

t is a time trend representing technical change;

the βs are unknown parameters to be estimated;

the v_{it}s are random errors, assumed to be i.i.d. and have N(0, σ_v^2)-distribution, independent of the u_{it}s; and

the u_{it}s are the technical inefficiency effects, with appropriately defined structure.

The above model has the time trend, t, interacted with the input variables. This allows for non-neutral technical change.

The technical efficiencies of each firm in each year can be predicted using the approach outlined in earlier chapters. That is, we obtain the conditional expectation of exp($-u_{it}$), given the value of $e_{it}=v_{it}-u_{it}$. Since u_{it} is a non-negative random variable, these technical efficiency predictions are between zero and one, with a value of one indicating full technical efficiency.

In this parametric case, we can use measures of technical efficiency and technical change to calculate the Malmquist TFP index via equations 11.4 to 11.6. The technical efficiency measures,

$$\mathrm{TE}_{it}=\mathrm{E}(\exp(-u_{it})|e_{it}), \tag{11.14}$$

where $e_{it}=v_{it} - u_{it}$, can be used to calculate the efficiency change component. That is, by observing that $d_o^t(\mathbf{x}_{it},y_{it})=\mathrm{TE}_{it}$ and $d_o^s(\mathbf{x}_{is},y_{is})=\mathrm{TE}_{is}$ we calculate the efficiency change index as:

$$\text{Efficiency change} = \mathrm{TE}_{it}/\mathrm{TE}_{is}. \tag{11.15}$$

This measure can be compared directly to equation 11.5.

The technical change index between period s and t for the i-th firm can be calculated directly from the estimated parameters. One first evaluates the partial derivatives of the production function with respect to time using the data for the i-th firm in periods s and t. Then the technical change index between the adjacent periods s and t is calculated as the geometric mean of these two partial derivatives. When a translog function is involved, this is equivalent to the exponential of the arithmetic mean of the log derivatives. That is,

$$\text{Technical change} = \exp\left\{\frac{1}{2}\left[\frac{\partial \ln y_{is}}{\partial s}+\frac{\partial \ln y_{it}}{\partial t}\right]\right\}. \tag{11.16}$$

This measure may be compared directly with equation 11.6. The indices of technical efficiency change and technical change obtained using equations 11.15 and 11.16 can then be multiplied together to obtain a Malmquist TFP index, as defined in equation 11.4.

Some issues are worth noting. First, the above technical change measure involves derivative calculations, which appears to contradict the earlier comments that these indices are derived from distance measures. It can be easily shown (for the translog case in which a time trend is used to represent technical change) that the geometric mean of the distance ratios in equation 11.6 are equivalent to the geometric means of the derivative measures.[5]

One possible criticism of the above method is that, if scale economies are important, the TFP index may produce biased measures because the productivity changes due to scale changes are not captured. One possible solution to this problem is to impose CRS upon the estimated production technology.[6] Another option is to use the approach proposed by Orea (2002), who uses Diewert's quadratic identity to derive a Malmquist TFP decomposition identical to that proposed above, and then suggests that the scale issue can be addressed in a manner similar to that used by Denny, Fuss and Waverman (1981). This involves the inclusion of a scale change component to the TFP measure,

$$\text{Scale change} = \exp\left\{\frac{1}{2}\sum_{n=1}^{N}\left[\varepsilon_{nis}SF_{is} + \varepsilon_{nit}SF_{it}\right]\ln(x_{nit}/x_{nis})\right\}, \qquad (11.17)$$

where $SF_{is} = (\varepsilon_{is} - 1)/\varepsilon_{is}$, $\varepsilon_{is} = \sum_{n=1}^{N}\varepsilon_{nis}$ and $\varepsilon_{nis} = \dfrac{\partial \ln q_{is}}{\partial \ln x_{nis}}$.

This scale change index is equal to one if the production technology is CRS. That is, when the scale elasticity (ε_{is}) equals one. For more information on this measure, see Orea (2002).

11.5 An Empirical Application

This empirical analysis utilises the rice industry data (that was used in earlier Chapters) to construct indices of TFP growth using the two methods described in Section 11.3. The sample data comprise annual measures of output and three inputs (land, labour, and fertiliser) for 43 rice farmers from the Philippines over the years from 1990 to 1997.

In the stochastic frontier approach, we specify a translog stochastic frontier production function for these rice farmers, with non-neutral technical change, as in

[5] To see this, refer to Fuentes et al (2001) for an expression for a translog technical change measure obtained directly from a ratio of translog distance measures, which utilise the period t technology. If one then also specifies the corresponding technical change measure derived from the period t+1 technology, and then finds the geometric mean of the two measures, the resulting expression will be equivalent to the expression obtained using equation 11.16.

[6] For example, see Nishimizu and Page (1982).

equation 11.13. The u_{it} are assumed to be i.i.d. $N^{+}(0, \sigma_U^2)$ random variables. Thus, the model does not assume that technical inefficiency is time invariant or that it follows a particular parametric structure. This permits the greatest degree of flexibility in the possible patterns of technical efficiency, both for a particular firm over time and among firms. However, if in reality a more restrictive parametric structure was the "truth", then the approach used here does not use that information and, hence, produces econometric estimators with larger variances.

The maximum-likelihood estimates of the parameters of the translog stochastic frontier model are obtained using the computer program, FRONTIER Version 4.1. These estimates are presented in Table 11.3. Asymptotic standard errors are presented beside each estimate. Also note that the data were mean corrected prior to estimation. Hence, the first-order parameters are interpreted as the elasticities at the sample means.

Table 11.3 Maximum-Likelihood Estimates of the Stochastic Frontier Model

Coefficient	Estimate	Standard Error	t-ratio
β_0	0.342	0.033	10.230
β_1	0.453	0.063	7.223
β_2	0.286	0.062	4.623
β_3	0.232	0.036	6.391
β_t	0.015	0.007	2.108
β_{11}	-0.509	0.225	-2.263
β_{12}	0.613	0.169	3.622
β_{13}	0.068	0.144	0.475
β_{1t}	0.005	0.024	0.215
β_{22}	-0.539	0.264	-2.047
β_{23}	-0.159	0.148	-1.073
β_{2t}	0.024	0.026	0.942
β_{33}	0.021	0.093	0.230
β_{3t}	-0.034	0.018	-1.893
β_{tt}	0.015	0.007	2.176
σ^2_s	0.223	0.025	9.033
γ	0.896	0.033	27.237
Log-likelihood	-70.592		

The estimated first-order coefficients are of similar magnitudes to those reported in Table 9.4 in Chapter 9. The only difference between the model specification in Table 9.4 and that used here is that the current model includes a time-squared variable, to allow for non-monotonic technical change, plus time interacted with

each (log) input variable, to allow for non-neutral technical change. An LR test indicates that these included variables are not significant additions at the 5% level of significance, but are significant at the 10% level. Given that we have a number of observations and, hence, reasonable degrees of freedom, we retain the more general model, especially given that our focus is on technical change and related measures in this empirical exercise.

As was noted in the analyses of these data in earlier chapters, the elasticity associated with land is the largest. The sum of the three production elasticities (0.45+0.29+0.23) is 0.97, suggesting very mild decreasing returns to scale at the sample mean data point. The coefficient of time is 0.015, which indicates mean technical progress of 1.5% per year. The coefficient of time squared is positive and significant (at the 5% level), indicating that the rate of technical change increases at an increasing rate through time. The coefficients of time interacted with the land, labour and fertiliser input variables are near zero, positive and negative, respectively, suggesting that technical change has been labour-saving but fertiliser-using over this period. Visually, this indicates that the isoquant is shifting inwards at a faster rate over time in the labour-intensive part of the input space. This is most likely a consequence of the rising relative cost of labour as the process of development continues in the Philippines.

Annual percentage change measures of technical efficiency change (TEC), technical change (TC), scale change (SC) and total factor productivity change (TFPC) were calculated for each firm in each pair of adjacent years using the methods described in Section 11.3.2. These measures have been averaged across firms and then converted into cumulative percentage change measures, which are reported in Table 11.4 and plotted in Figure 11.2.

Table 11.4 Cumulative Percentage Change Measures of TEC, TC, SC and TFPC, using SFA

Year	TEC	TC	SC	TFPC
1990	0.0	0.0	0.0	0.0
1991	-0.6	-2.7	0.0	-3.3
1992	13.2	-4.2	-0.3	8.7
1993	9.9	-4.0	0.1	6.0
1994	1.6	-2.1	-0.2	-0.8
1995	7.9	0.8	0.4	9.1
1996	-12.9	4.9	0.7	-7.4
1997	8.1	10.2	0.7	19.1

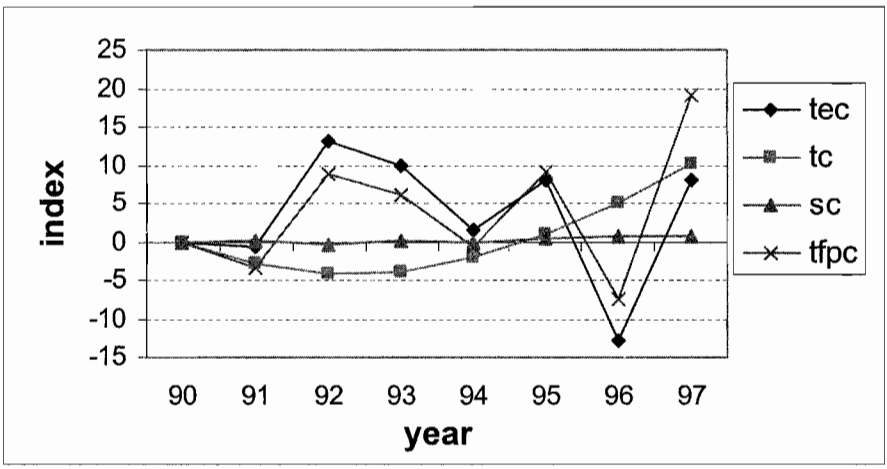

Figure 11.2 Cumulative Percentage Change Measures of TEC, TC, SC and TFPC using SFA

A number of points can be made regarding the results in Table 11.4 and Figure 11.2. First, we note that TFPC is quite stochastic. This is primarily due to the effect of stochastic TEC, which is most likely a consequence of climatic factors, such as rainfall and temperature, affecting rice yields. This stochastic behaviour can be an important factor in analyses of agricultural industries that are exposed to the elements, such as rice farming, but are less of an issue in industrial applications, such as motor vehicle manufacturing. In situations such as the rice farming case considered here, many agricultural economists do not only report the TFPC measure of 19.1 %, but would also report some kind of smoothed measure.

This could be obtained, for example, by running a regression of TFP upon a time trend. When this is done with the data in the final column of Table 11.4 we obtain an estimated time-trend coefficient of 1.5, indicating average annual TFP change of 1.5% per year. This is significantly less than the value of 19.1/8 = 2.4% that is obtained directly from the results. Alternatively, if we had finished our data series in 1996 instead of 1997, the directly calculated TFP measure would have been the value, –7.4/7 = –1.1%. This illustrates the effect of finishing or starting the analysis in a good or bad season, and the importance of using smoothing in these cases.

A similar issue to the above can also be found in manufacturing applications, where the productivity of the industry (for example, motor vehicle manufacturing) is likely to be affected by the current state of the macro-economic business cycle, via the effect of consumer demand factors upon capacity utilisation. To attempt to correct this issue, a number of analysts report TFP measures measured from peak to peak, based on the assumption that capacity utilisation rates are similar in boom periods.

Putting these issues of climate and capacity utilisation aside for now, from the bottom line of Table 11.4 we see that over this eight-year period, TFP increased by 19.1%, due to the 10.2% upward shift in the technology, the 8.1% increase in TE and the small 0.7% increase in productivity due to scale effects. The small size of the scale effect is not surprising, given that the estimated technology had scale economies close to one, indicating that the function was approximately CRS at the sample means, plus there were minimal changes in average farm sizes over this period. It is also clear that TC is increasing at an increasing rate, as was suggested in our earlier discussion of the signs of the estimated coefficients.

The above discussion of sample average patterns is of considerable interest, but it is based upon aggregate results. It should not be forgotten that a Malmquist index analysis utilising panel data produces a rich quantity of information on TEC, TC, SC and TFPC for each firm between each pair of adjacent time periods. These measures are not reported in this Chapter to conserve space, but are available from the authors via the CEPA web site (www.uq.edu.au/economics/cepa/crob2005).

DEA Results

The same sample data were used to calculate indices of TEC, TC, SC and TFPC using the DEA-like methods described in Section 11.3.1. Calculations were done using the DEAP Version 2.1 computer program. These indices were converted into percentage change measures to allow ease of comparison with the SFA results. These DEA results are summarised in Table 11.5 and Figure 11.3.

The DEA results differ from the SFA results in a number of aspects. First the DEA TFPC measures are more stochastic than the SFA measures. This is not surprising, given that the DEA method involves the calculation of a separate frontier in each year, while the SFA method uses all eight years of data to estimate the frontiers for all eight years, with "smooth" changes in the frontier allowed via the time trend specification of technical change. The net effect of this is less dramatic changes in frontier shapes and hence in shadow shares and, hence, in TFPC measures, in the SFA case.[7]

Another point of difference is that the stochastic nature of the DEA TFPC is more due to TC, in this case. This is essentially due to the above-mentioned year-to-year flexibility of the DEA method. When all farms face a bad year in terms of rainfall, the DEA method tends to interpret this as technical regress (i.e. negative TC) while the SFA method interprets it as a decline in TE.[8]

---

[7] The effect of shadow shares on TFPC measures is discussed shortly.

[8] Nghiem and Coelli (2002) face a similar problem in an analysis of TFP change in Vietnamese rice production. They develop two different amended Malmquist DEA techniques to address the issue. The first is a three-year window method that uses data from periods t, t-1 and t-2 to construct the period t frontier, while the second method involves the use of data from period t plus all earlier years of data available to construct the period t frontier. This latter "cumulative" method has the additional property that it does not permit technological regress to be estimated.

Table 11.5 Cumulative Percentage Change Measures of TEC, TC, SC and TFPC using DEA

Year	TEC	TC	SC	TFPC
1990	0.0	0.0	0.0	0.0
1991	6.6	-16.1	8.4	-3.0
1992	20.1	-0.1	0.0	17.6
1993	15.2	2.0	2.0	16.6
1994	7.3	-8.9	6.4	2.2
1995	11.7	-6.0	8.3	11.7
1996	8.6	-19.1	5.3	-6.6
1997	6.8	48.9	-7.9	36.6

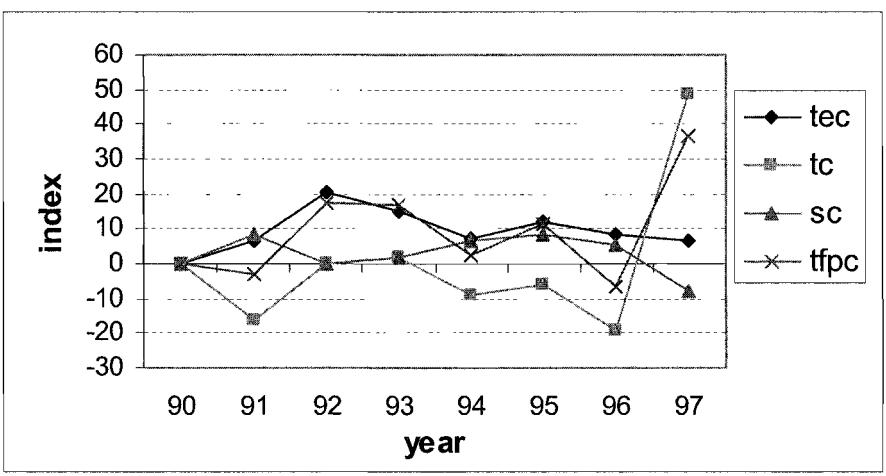

Figure 11.3 Cumulative Percentage Change Measures of TEC, TC, SC and TFPC using DEA

Overall, the two TFPC patterns are similar, but the DEA results indicate a higher overall TFP growth. To obtain a better idea of how the two TFPC measures compare, we plot them on the one graph in Figure 11.4, along with a Tornqvist price-based index number (PIN) TFPC measure calculated using the same data.

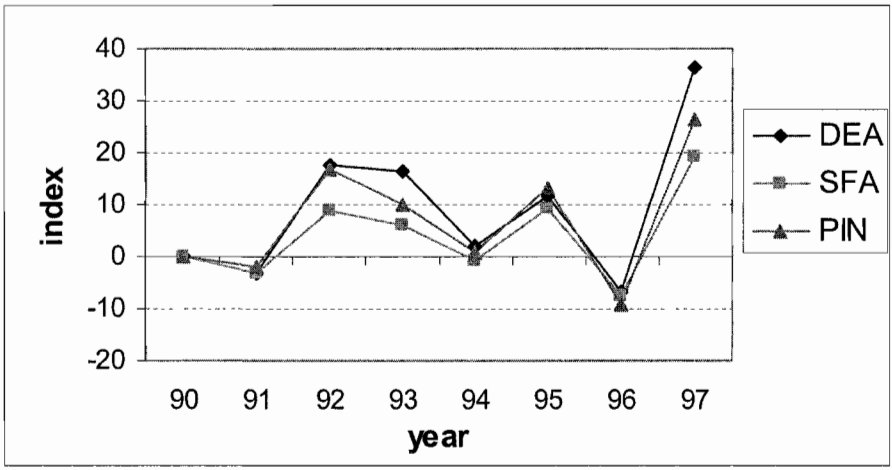

Figure 11.4 Cumulative TFP Change using DEA, SFA and PIN

We observe a number of things in Figure 11.4. First, the three measures follow a very similar pattern over the eight-year period, which is reassuring. Second, the DEA TFPC measure is the most volatile, most likely for the reasons discussed above. Third, the aggregate measures of TFPC range from 36.6% for DEA, down to 26.2% for PIN and then 19.1% for SFA. The majority of these observed differences can be put down to the effect of differences in the input shares that are used to weight the input changes in the TFPC calculations. This comment may come as a surprise to some researchers who may have seen it often written that these Malmquist index methods (which utilise panel data) do not need price information. This is true – they do not need *explicit* price information (e.g. market prices). However, as pointed out by Coelli and Rao (2001), even though the DEA and SFA methods do not utilise explicit price information, they do utilise *implicit* price information (i.e. shadow price information) derived from the shape of the estimated frontiers. Thus, unless the different methods use identical (implicit or explicit) input share measures, the TFPC measures derived from them will most likely differ to some degree.

To illustrate this issue using the present empirical application, we present sample average input share measures for the SFA, DEA and PIN methods in Table 11.6. The DEA shadow shares have been calculated the using the methods described in Coelli and Rao (2001).[9] The SFA shadow shares are obtained by scaling the estimated production elasticities at each data point by the sum of the elasticities. This ensures that they sum to one at each data point.[10] Finally, the PIN input shares are calculated in the traditional way using the observed price data.

[9] This involves the use of the DEA shadow price weights discussed in Chapter 6.

[10] The logic behind this measure can be seen when one looks closely at the scale change (SC) measure presented in equation 11.17. When this SC measure is added to the TFPC measure it has the effect of changing the TFPC measure from one that uses the raw elasticities as input weights (which need not add

Table 11.6 Sample Average Input Shares

	Land	Labour	Fertiliser
DEA	0.388	0.358	0.254
SFA	0.478	0.334	0.188
PIN	0.487	0.381	0.131

From the values presented in Table 11.6, it is evident that some inputs shares differ substantively among the three methods. When it is also noted from the sample data that the amount of labour used per hectare reduced over this period, it is not surprising that SFA has the lowest calculated TFPC measure because it has the highest ratio of land share to labour share, while DEA has the highest calculated TFPC measure because it has the lowest ratio of land share to labour share.

The above discussion helps explain why the TFPC measures differ across the methods. However, it does not give us much guidance as to which measure or measures are "correct". The answer to this question is rarely straight forward, and tends to differ according to the industry being studied, the quality of the data and the purpose for which the analyst wishes to use the TFPC measures obtained. In this particular industry the data is quite noisy, so we would tend to prefer SFA over DEA. Also, since this is an industry in which there are limited market distortions (e.g. regulatory factors), one may expect that market prices and shadow prices should be roughly similar. Given these two observations, and the fact that the TFPC measures (and input share measures) in SFA and PIN are similar, one could conclude that a reasonable measure of TFPC in this industry is between 19% and 26% over this eight-year period, which is equivalent to an average annual change of between 2.4% and 3.2% per year.

One additional point worth noting is to emphasise that the reported industry-level measures of TFPC and its components are unweighted averages of the firm-level measures. In some instances, when one suspects that TFPC could be correlated with firm size, a weighted average may provide a more accurate estimate of the industry-level TFPC measure. In some instances, the difference between weighted and unweighted measures can be substantial. For example, in an analysis of TFPC in Australian electricity generation, Coelli (2002) found that the SFA TFPC measure increased from –0.7% per year to 0% per year, while the DEA measure increased from 0.7% per year to 1.2% per year.

to one) to one that uses these scaled elasticities instead, and hence accommodates the possible effects of scale changes upon TFPC.

11.6 Conclusions

We conclude this chapter by summarising some of the relative merits of using frontier approaches (such as SFA and DEA) to conduct Malmquist TFP calculations versus using the traditional PIN approaches (such as Tornqvist or Fisher indices). Some of the advantages of the frontier approach are:

- The frontier approach does not require price information;
- It does not assume all firms are fully efficient;
- It does not need to assume a behavioural objective such as cost minimisation or revenue maximisation;[11]
- It permits TFP change (TFPC) to be decomposed into components, such as technical change (TC), technical efficiency change (TEC) and scale change (SC).

However, an important advantage of the Tornqvist approach is that it can be calculated using only two data points, while the frontier approach needs a number of firms to be observed in each time period so that the frontier technology in each year can be estimated. Thus, if one has suitable panel data, the frontier approach provides richer information and makes fewer assumptions. However, if only aggregate time-series data are available, then the Tornqvist approach allows one to obtain useful estimates of TFP change, given that the above-listed assumptions are reasonable.

[11] These assumptions are used in constructing an economic theory justification for the Törnqvist and Fisher TFP index numbers. Alternatively, if one was instead using an axiomatic justification, one could argue that these assumptions are not required.

12. CONCLUSIONS

12.1 Summary of Methods

Scattered throughout earlier chapters are a number of discussions of the characteristics and relative merits of the various methods that we have considered. The purpose of this final chapter is to bring together some of this material so we can reflect upon it.

We have considered four principal methods:

1. least-squares (LS) econometric production models,
2. total factor productivity (TFP) indices (Tornqvist/Fisher),
3. data envelopment analysis (DEA), and
4. stochastic frontiers (SF).

These four methods differ in various ways. For example, some are parametric while others non-parametric. Some can accommodate the effects of data noise while others cannot. Some but not all can be used to measure technical efficiency and allocative efficiency. Some can be applied using time series data while others cannot. Some methods do not require price data. These and other issues are summarised in Table 11.1 below.

Table 12.1 Summary of the Properties of the Four Principal Methods

Attribute	LS	TFP	DEA	SF
Parametric method	yes	no	no	yes
Accounts for noise	yes	no	no	yes
Can be used to measure:				
technical efficiency	no	no	yes	yes
allocative efficiency	yes	no	yes	yes
technical change	yes	no	yes	yes
scale effects	yes	no	yes	yes
TFP change	yes	yes	yes	yes
Data used:				
cross sectional	yes	yes	yes	yes
time series	yes	yes	no	no
panel	yes	yes	yes	yes
Basic method requires data on:[1]				
input quantities	yes	yes	yes	yes
output quantities	yes	yes	yes	yes
input prices	no	yes	no	no
output prices	no	yes	no	no

12.2 Relative Merits of the Methods

Efficiency is generally measured using either DEA or stochastic frontier methods. Some of the advantages of stochastic frontiers over DEA are:

- it accounts for noise, and
- it can be used to conduct conventional tests of hypotheses.

while some disadvantages are:

- the need to specify a distributional form for the inefficiency term, and
- the need to specify a functional form for the production function (or cost function, etc.),

Technological change (or TFP) is usually measured using either least squares econometric methods or Tornqvist/Fisher index numbers. Some of the advantages of index numbers over least-squares econometric methods are:

[1] Note that this applies to the basic primal method only. In situations where cost or profit functions are estimated these requirements differ. For example, a stochastic production function requires data on input and output quantities, while a (long run) stochastic cost frontier requires data on total cost, output quantities and input prices.

- only two observations are needed,
- they are easy to calculate, and
- the method does not assume a smooth pattern of technical progress,

while the principal disadvantage is:

- it requires both price and quantity information.

Both of these approaches assume that firms are technically efficient (which is unlikely to be true). To relax this assumption one can use frontier methods (assuming panel data are available) to calculate TFP change. Some of the advantages of this frontier approach over the Tornqvist/Fisher index numbers approach are that:

- it does not require price information,
- it does not assume all firms are fully efficient,
- it does not require the assumption of cost minimisation and revenue maximisation, and
- it permits TFP to be decomposed into technical change and technical efficiency change.

However, an important advantage of the index-number approach is that it:

- only requires two data points, say observations on two firms in one time period or observations on one firm in two time periods, while the frontier approaches need a number of firms to be observed in each time period so that the frontier technology in each year can be calculated.[2]

12.3 Some Final Points

We now have a collection of very powerful and flexible tools at our disposal, to help in our analyses of efficiency and productivity. But before concluding this book we should make mention of some of the many pitfalls that a performance measurement analysis may suffer from.

- Treating inputs and/or outputs as homogenous commodities when they are heterogenous may bias results.
- There may be measurement error in the data.

[2] Note that the DEA approach can be used if only one observation in each year is available. However, in this case one must assume there is no inefficiency, or alternatively include past observations in the calculation of period-t technology. See Grosskopf (1993, p.182) for more on this.

- Exclusion of an important input or output may bias the results.
- Not accounting for environmental differences (both physical and regulatory) may give misleading results.
- Most methodologies do not account for multi-period optimisation or risk in management decision making.

The above points relate to all the methods considered in this book. For frontier methods, in particular, we add the points:

- The efficiency scores are only relative to the best firms in the sample. The inclusion of extra firms (say from another country) may reduce efficiency scores.
- Be careful when comparing the mean efficiency scores from two studies. They only reflect the dispersion of efficiencies within each sample. They say nothing about the efficiency of one sample relative to the other.
- Measurement error and other noise may influence the shape and position of the frontier.
- Outliers may influence results.

To illustrate the importance of some of the above points, we pose a few questions. The following questions assume that a researcher has conducted a quick analysis of some data and is at the stage of trying to interpret the preliminary results obtained.

Q1) The researcher observes that a particular firm has lower productivity relative to other firms in an industry. Why may this be so? It could be due to one or more of:

- technical (managerial) inefficiency,
- scale inefficiency,
- omitted variables,
- quality differences in inputs and outputs,
- measurement error,
- unused capacity due to lumpy investment,
- environment:
    - physical and/or
    - regulatory.

Q2) The researcher observes that the TFP of a firm has improved from one period to the next. Why may this be so? It could be due to one or more of:

- improved technical efficiency,
- technical progress,
- scale improvements,
- changes in quality of inputs and/or outputs,
- measurement error,
- changes in environment,
- utilisation of idle capacity.

Q3) The researcher observes that the unit cost for a firm has declined from one period to the next. Why may this be so? It could be due to one or more of:[3]

- all in above list, plus:
- increased allocative efficiency (in input choices),
- favourable input price changes (need to select price indices carefully).[4]

Q4) The researcher observes that the profit for a firm has increased from one period to the next. Why may this be so? It could be due to one or more of:

- all in above list, plus:
- increased allocative efficiency (in output choices),
- favourable output price changes (need to select price indices carefully),
- the firm has simply become larger.

Although "performance" is a somewhat slippery concept, with careful attention to the issues listed above, the performance measurement tools described in this book can provide valuable information in many situations.

[3] Note that we are assuming no "creative" accounting methods are used.

[4] For example, if money values are not deflated or if you use an inappropriate deflator (e.g., the CPI), the cost figures may reflect more than improved performance.

APPENDIX 1: COMPUTER SOFTWARE

In this appendix we provide details on the computer software that is used in this book. Five computer programs are used:

1. SHAZAM - a general purpose econometrics package
2. LIMDEP - a general purpose econometrics package
3. DEAP - a data envelopment analysis (computer) program (Coelli, 1996b).
4. FRONTIER - a computer program for the estimation of stochastic frontier models (Coelli, 1996a).
5. TFPIP - a total factor productivity index (computer) program written by Tim Coelli.

The SHAZAM and LIMDEP computer programs are a widely used econometrics software packages. They can be used to estimate a large number of econometric models. For further information on these computer programs, including information on how to purchase them, refer to the web sites:

http://shazam.econ.ubc.ca/

and

http://www.limdep.com/

The remaining three computer programs (listed above) were written by Tim Coelli, specifically for the measurement of efficiency and/or productivity. Information on these three computer programs can be obtained from the Centre for Efficiency and Productivity Analysis (CEPA) web site:

http://www.uq.edu.au/economics/cepa

where copies of these programs (including manuals) may be downloaded free of charge. We now discuss the use of these latter three computer programs.

DEAP Version 2.1: A Data Envelopment Analysis (Computer) Program

This computer program has been written to conduct data envelopment analyses (DEA). The computer program can consider a variety of models. The three principal options are:

1. Standard CRS and VRS DEA models that involve the calculation of technical and scale efficiencies (where applicable). These methods are outlined in Chapter 6.
2. The extension of the above models to account for cost and allocative efficiencies. These methods are outlined in Section 7.2.
3. The application of Malmquist DEA methods to panel data to calculate indices of total factor productivity (TFP) change; technological change; technical efficiency change and scale efficiency change. These methods are discussed in Chapter 10.

All methods are available in either an input or an output orientation (with the exception of the cost efficiencies option). The output from the program includes, where applicable, technical, scale, allocative and cost efficiency estimates; slacks; peers; targets; TFP and technological change indices.

The DEAP computer program is written in Fortran (Lahey F77LEM/32) for IBM compatible PCs. It is a DOS program but can be easily run from WINDOWS using WINDOWS EXPLORER. The program involves a simple batch file system where the user creates a data file and a small file containing instructions. The user then starts the program by typing "DEAP" at the DOS prompt[1] and is then prompted for the name of the instruction file. The program then executes these instructions and produces an output file which can be read using a text editor, such as NOTEPAD, or any program that can accept text files, such as WORD or EXCEL.

The execution of DEAP Version 2.1 on PC generally involves five files:

1. The executable file, DEAP.EXE
2. The start-up file, DEAP.000
3. A data file (for example, called TEST-DTA.TXT)
4. An instruction file (for example, called TEST-INS.TXT)
5. An output file (for example, called TEST-OUT.TXT).

[1] The program can also be run by double-clicking on the DEAP.EXE file in WINDOWS EXPLORER. The use of WINDOWS EXPLORER is discussed at the end of this appendix.

The executable file and the start-up file is supplied on the disk. The start-up file, DEAP.000, is a file that stores key parameter values that the user may or may not need to alter.[2] The data and instruction files must be created by the user prior to execution. The output file is created by DEAP during execution. Examples of data, instruction and output files are listed in Chapters 6 and 7.

Data file

The program requires that the data be listed in a text file[3] and expects the data to appear in a particular order. The data must be listed by observation (i.e., one row for each firm). There must be a column for each output and each input, with all outputs listed first and then all inputs listed (from left to right across the file). For example, for 40 observations on two outputs and two inputs there would be four columns of data (each of length 40) listed in the order: y1, y2, x1, x2.

The cost efficiencies option requires that price information be supplied for the inputs. These price columns must be listed to the right of the input data columns and appear in the same order. That is, for three outputs and two inputs, the order for the columns must be: y1, y2, y3, x1, x2, w1, w2, where w1 and w2 are input prices corresponding to input quantities, x1 and x2.

The Malmquist option is used with panel data. For example, for 30 firms observed in each of 4 years, all data for year 1 must be listed first, followed by the year 2 data listed underneath in the same order (of firms) and so on. Note that the panel must be "balanced", i.e., all firms must be observed in all time periods.

A data file can be produced using any number of computer packages. For example:

- using a text editor (such as NOTEPAD),
- using a word processor (such as WORD) and saving the file in text format,
- using a spreadsheet (such as EXCEL) and printing to a file, or
- using a statistics package (such as SHAZAM or LIMDEP) and writing data to a file.

Note that the data file should only contain numbers separated by spaces or tabs. It should not contain any column headings.

[2] At present this file only contains two parameters. One is the value of a variable (EPS) used to test inequalities with zero and the other is a flag that can be used to suppress the printing of the firm-by-firm reports in the output file. This text file may be edited if the user wishes to alter this value.
[3] All data, instruction and output files are (ASCII) text files.

Instruction file

The instruction file is a text file that is usually constructed using a text editor or a word processor. The easiest way to create a new instruction file is to edit one of the example instruction files that are supplied with the program and then save the edited file under a different file name. The best way to describe the structure of the instruction file is via examples. Refer to the examples in Chapters 6 and 7.

Output file

As noted earlier, the output file is a text file that is produced by DEAP when an instruction file is executed. The output file can be read using a text editor, such as NOTEPAD, or using a word processor, such as WORD. The output may also be imported into a spreadsheet program, such as EXCEL, to allow further manipulation into tables and graphs for subsequent inclusion into report documents.

FRONTIER Version 4.1: A Computer Program for Stochastic Frontier Estimation

The FRONTIER computer program is very similar in construction to the DEAP computer program. It has been written to provide maximum-likelihood estimates of the parameters of a number of stochastic frontier production and cost functions. The stochastic frontier models considered can accommodate (unbalanced) panel data and assume firm effects that are distributed as truncated normal random variables. The two primary model specifications considered in the program are:

1. The Battese and Coelli (1992) time-varying inefficiencies specification, which is discussed in Section 10.5.
2. The Battese and Coelli (1995) model specification in which the inefficiency effects are directly influenced by a number of variables. This model is discussed in Section 10.6.

The computer program also permits the estimation of other models that have appeared in the literature through the imposition of simple restrictions. Estimates of standard errors are also calculated, along with individual and mean efficiency estimates.

The program can accommodate cross-sectional and panel data; time-varying and time-invariant inefficiency effects; cost and production functions; half-normal and truncated normal distributions; and functional forms which have a dependent variable in logged or original units.

The execution of FRONTIER Version 4.1 on an IBM PC generally involves five files:

1. The executable file, FRONT41.EXE
2. The start-up file, FRONT41.000
3. A data file (for example, called TEST-DTA.TXT)
4. An instruction file (for example, called TEST-INS.TXT)
5. An output file (for example, called TEST-OUT.TXT).

The start-up file, FRONT41.000, contains values for a number of key variables, such as the convergence criterion, printing flags and so on. This text file may be edited if the user wishes to alter any values. The data and instruction files must be created by the user prior to execution.[4] The output file is created by FRONTIER during execution. Examples of data, instruction and output files are presented in Chapter 9.

The program requires that the data be stored in an text file and is quite particular about the order in which the data are listed. Each row of data should represent an observation. The columns must be presented in the following order:

1. firm number (an integer in the range 1 to N);
2. period number (an integer in the range 1 to T);
3. dependent variable;
4. regressor variables; and
5. variables influencing the inefficiency effects (if applicable).

The observations can be listed in any order but the columns must be in the stated order. There must be at least one observation on each of the N firms and there must be at least one observation in time period 1 and in time period T. If cross-sectional data are involved, then column 2 (the time-period column) must contain the value "1" throughout. Note that the data must be suitably transformed if a functional form other than a linear function is required. The Cobb-Douglas and translog functional forms are the most often used functional forms in stochastic frontier analyses. Examples involving the translog form are provided in Chapters 9 and 10.

The program can receive instructions either from a file or directly from the keyboard. After typing "FRONT41" to begin execution, the user is asked whether instructions will come from a file or the terminal. An example of an instruction file is listed in Chapters 9 (Table 9.2). If the interactive (terminal) option is selected, questions will be asked in the same order as they appear in the instruction file.

---

[4]Note that a model can be estimated without an instruction file if the program is used interactively.

The Three-step Estimation Method

The program follows a three-step procedure in estimating the maximum-likelihood estimates of the parameters of a stochastic frontier production function.[5] The three steps are:

1. Ordinary least-squares (OLS) estimates of the parameters of the function are obtained. All β-estimators with the exception of the intercept, β_0, will be unbiased.
2. A two-phase grid search of γ is conducted, with the β parameters (excepting β_0) set to the OLS values and the β_0 and σ^2 parameters adjusted according to the corrected ordinary least-squares formula presented in Coelli (1995c). Any other parameters (μ, η or δs) are set to zero in this grid search.
3. The values selected in the grid search are used as starting values in an iterative procedure (using the Davidon-Fletcher-Powell Quasi-Newton method) to obtain the final maximum-likelihood estimates.

Program Output

The ordinary least-squares estimates, the estimates after the grid search and the final maximum-likelihood estimates are all presented in the output file. Approximate standard errors are taken from the direction matrix used in the final iteration of the Davidon-Fletcher-Powell procedure. This estimate of the covariance matrix is also listed in the output.

Estimates of individual technical or cost efficiencies are calculated using the expressions presented in Battese and Coelli (1992, 1993). When any estimates of mean efficiencies are reported, these are simply the arithmetic averages of the individual efficiencies.

[5]If starting values are specified in the instruction file, the program will skip the first two steps of the procedure.

TFPIP Version 1.0: A Total Factor Productivity Index (Computer) Program

TFPIP is a simple computer program that can be used to calculate Fisher and Tornqvist TFP indices (both regular and transitive). Input and output quantity indices are also reported in the output produced by the program. Refer to Chapter 4 for further details on these index numbers.

The TFPIP computer program is structured in a similar manner to the FRONTIER and DEAP computer programs. Execution generally involves four files:

1. The executable file TFPIP.EXE
2. A data file (for example, called TEST-DTA.TXT)
3. An instruction file (for example, called TEST-INS.TXT)
4. An output file (for example, called TEST-OUT.TXT).

Examples of data, instruction and output files are listed in Chapter 4.

The program requires that the data be listed in a text file and expects the data to appear in a particular order. The data must be listed by observation (i.e., one row for each firm). There must be a column for each output and input quantity and price. The data columns are listed as follows:

1. output quantities;
2. input quantities;
3. output prices; and
4. input prices.

The price columns should appear in the same order as the quantity columns. For example, for 40 observations on two outputs and three inputs, then there would be 10 columns of data (each of length 40) listed in the order: y1, y2, x1, x2, x3, p1, p2, w1, w2, w3.

Tips on using DEAP, FRONTIER or TFPIP in WINDOWS EXPLORER:

The DEAP, FRONTIER and TFPIP computer programs are all DOS programs. However, they can be easily manipulated using WINDOWS EXPLORER. The following steps illustrate how these programs can be used without a knowledge of DOS.

1. Within WINDOWS EXPLORER, use the FILE/NEW/FOLDER menu items to create a DEAP directory on the hard drive of the computer being used.
2. Download the DEAP zip file from the CEPA web site and save it in the DEAP directory.
3. Double-click the zip file to extract the DEAP computer program and the associated files.
4. Double click on some of the example data, instruction and output files to see their contents (e.g. view using NOTEPAD).
5. To practice executing DEAP, double-click on the DEAP.EXE file name. The program then asks for an instruction file name. Type in EG1-INS.TXT (and hit the RETURN key). DEAP will only take a few seconds to complete this small DEA example. To look at the output file (EG1-OUT.TXT), simply double-click on the EG1-OUT.TXT file name.

APPENDIX 2: PHILIPPINES RICE DATA

The International Rice Research Institute (IRRI) supplied data collected from 43 smallholder rice producers in the Tarlac region of the Philippines between 1990 and 1997. Details of the survey can be found in Pandey et al (1999). The data were used to construct observations on the variables listed in Table 5.1.

The PRICE, PROD, AREA, LABOR, NPK, AGE, EDYRS, HHSIZE, NADULT and BANRAT variables are exactly as reported in the file supplied by IRRI.

The input variable OTHER was constructed as a Laspeyres quantity index that combines inputs of seed, insecticides, herbicides and animals and tractors used during land preparation. The Laspeyres index formula was used instead of the more commonly-used Törnqvist formula because many farmers used only a subset of the inputs in this group – this would have forced the associated Tornqvist index numbers to zero. We chose Firm 17 in 1991 as the reference observation for the Lasypeyres index because in 1991 this firm used a nonzero amount of every input to produce a level of output that was in the neighbourhood of the median output of all firms in all years.

The AREAP series was constructed using predictions from a linear regression model developed by Fujimoto (1996). The Fujimoto model explains rice land rentals in ten villages in Indonesia, Malaysia, Thailand and the Philippines as a function of z_1 = average yield (kg/ha), z_2 = population pressure (persons/ha) and z_3 = the number of kinship-based tenancy contracts as a proportion of total tenancy contracts. To generate AREAP we set $z_1 = 1000 \times$ PROD/AREA, $z_2 = 3.09$ and $z_3 =$ 83.3 percent (these last two values were estimates based on IRRI data and statistics reported by Fujimoto).

The LABORP series was constructed as a quantity-weighted average of the implicit price of hired labour and a wage variable contained in the IRRI data file, using man-days of hired and family labour as weights.

The NPKP series was constructed by dividing the cost of fertiliser by NPK. The OTHERP series was constructed by dividing the cost of all other inputs by OTHER. All prices are in nominal terms.

The data is stored in the file RICE.CSV (this is a comma-delimited format that can be read straight into SHAZAM and EViews). Summary statistics are reported in Table 5.2

Table 5.1 Variable Descriptions

YEARDUM	Time Period
FMERCODE	Farmer Code
PROD	Output (tonnes of freshly threshed rice)
AREA	Area planted (hectares)
LABOR	Labour used (man-days of family and hired labour)
NPK	Fertiliser used (kg of active ingredients)
OTHER	Other inputs used (Laspeyres index = 100 for Firm 17 in 1991)
PRICE	Output Price (pesos per kg)
AREAP	Rental price of land (pesos per hectare)
LABORP	Labour price (pesos per hired man-day)
NPKP	Fertiliser price (pesos per kg of active ingredient)
OTHERP	Price of other inputs (implicit price index)
AGE	Age of household head (years)
EDYRS	Education of household head (years)
HHSIZE	Household size
NADULT	Number of adults in the household
BANRAT	Percentage of area classified as bantog (upland) fields

Table 5.2 Summary Statistics

VARIABLE	N	MEAN	ST.DEV	MINIMUM	MAXIMUM
YEARDUM	344	4.5000	2.2946	1.0000	8.0000
FMERCODE	344	22.000	12.428	1.0000	43.000
PROD	344	6.5403	5.1069	0.0900	31.100
AREA	344	2.1435	1.4580	0.2000	7.0000
LABOR	344	108.34	77.191	8.0000	437.00
NPK	344	189.23	169.80	10.000	1030.9
OTHER	344	125.34	158.24	1.4586	1083.4
PRICE	344	6.5313	1.5303	4.5000	9.0000
AREAP	344	5289.3	3505.7	313.87	27788.
LABORP	344	70.767	26.132	22.960	219.56
NPKP	344	14.885	3.4015	6.9029	32.396
OTHERP	344	32.836	24.038	3.3965	105.08
AGE	344	49.445	11.022	25.000	81.000
EDYRS	344	7.2442	1.9101	6.0000	14.000
HHSIZE	344	5.0262	2.0580	2.0000	14.000
NADULT	344	3.8488	1.8007	1.0000	10.000
BANRAT	344	0.7344	0.2933	0.0000	1.0000

REFERENCES

Abadir, K.M. and J.R. Magnus (2002), "Notation in Econometrics: A Proposal for a Standard", *Econometrics Journal*, 5, 76-90.

ABS (1989), "Development of Multifactor Productivity Estimates for Australia 1974-75 to 1987-88", *Australian Bureau of Statistics Information Paper* No. 5229.0, Canberra.

Afriat, S.N. (1972), "Efficiency Estimation of Production Functions", *International Economic Review*, 13, 568-598.

Aigner, D.J., and S.F. Chu (1968), "On Estimating the Industry Production Function", *American Economic Review*, 58, 826-839.

Aigner, D.J., C.A.K Lovell and P. Schmidt (1977), "Formulation and Estimation of Stochastic Frontier Production Function Models", *Journal of Econometrics*, 6, 21-37.

Ali, A.I., and L.M. Seiford (1993), "The Mathematical Programming Approach to Efficiency Analysis", in Fried, H.O., C.A.K. Lovell and S.S. Schmidt (Eds.), *The Measurement of Productive Efficiency: Techniques and Applications*, Oxford University Press, New York, 120-159.

Ali, M., and J.C. Flinn (1989), "Profit Efficiency Among Basmati Rice Producers in Pakistan Punjab", *American Journal of Agricultural Economics*, 71, 303-310.

Allen, R., A.D. Athanassopoulos, R.G. Dyson and E. Thanassoulis (1997), "Weight Restrictions and Value Judgements in DEA: Evolution, Development and Future Directions", *Annals of Operations Research*, 73, 13-34.

Allen, R.C., and W.E. Diewert (1981), "Direct versus Implicit Superlative Index Number Formulae", *Review of Economics and Statistics*, 63, 430-435.

Allen, R.G.D. (1975), *Index Numbers in Theory and Practice*, New York, Macmillan Press.

Althin, R. (1995), *Essays on the Measurement of Producer Performance*, Ph.D. Dissertation, Lund University, Lund.

Andersen, P., and N. Petersen (1993), "A Procedure for Ranking Efficient Units in Data Envelopment Analysis", *Management Science*, 39, 1261-1264.

Antle, J.M. (1984), "The Structure of U.S. Agricultural Technology, 1910-78", *American Journal of Agricultural Economics*, 66, 414-421.

Atkinson, S.E., and J.H. Dorfman (2005), "Bayesian Measurement of Productivity and Efficiency in the Presence of Undesirable Outputs: Crediting Electric Utilities for Reducing Air Pollution", *Journal of Econometrics*, 126, 445-468.

Atkinson. S.E., R. Färe and D. Primont (1998), "Stochastic Estimation of Firm Inefficiency Using Distance Functions", Working Paper, Department of Economics, University of Georgia, Athens, GA.

Atkinson, S.E., and D. Primont (1998), "Stochastic Estimation of Firm Technology, Inefficiency and Productivity Growth Using Shadow Cost and Distance Functions", Working Paper, Department of Economics, University of Georgia, Athens, GA.

Balk, B.M. (1995), "Axiomatic Price Index Theory: A Survey", International Statistical Review, 63, 69-93.

Balk, B.M. (1997), Industrial Price, Quantity, and Productivity Indices: Micro-Economic Theory", Mimeographed, Statistics Netherlands, 171.

Balk, B.M. (1998), *Industrial Price, Quantity, and Productivity Indices: The Micro-Economic Theory and an Application*, Kluwer Academic Publishers Boston.

Balk, B.M. (2001), "Scale Efficiency and Productivity Change", *Journal of Productivity Analysis*, 15, 159-183.

Balk, B.M. (2003), "On the Relationship between Gross-output and Value-added Based Productivity Measures: The Importance of the Domar Factor", *Working Paper 05/2003*, Centre for Applied Economic Research, University of New South Wales, Sydney.

Balk, B.M., and R. Althin (1996), "A New, Transitive Productivity Index", *Journal of Productivity Analysis*, 7, 19-27.

Banker, R.D. (1996), "Hypothesis Test using Data Envelopment Analysis", *Journal of Productivity Analysis*, 7, 139-160.

Banker, R.D., A. Charnes and W.W. Cooper (1984), "Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis", *Management Science*, 30, 1078-1092.

Banker, R.D., and R.C. Morey (1986a), "Efficiency Analysis for Exogenously Fixed Inputs and Outputs", *Operations Research*, 34, 513-521.

Banker, R.D., and R.C. Morey (1986b), "The Use of Categorical Variables in Data Envelopment Analysis", *Management Science*, 32, 1613-1627.

Banker, R.D., and R.M. Thrall (1992), "Estimation of Returns to Scale Using Data Envelopment Analysis", *European Journal of Operational Research* 62, 74-84.

Battese, G.E. (1992), "Frontier Production Functions and Technical Efficiency: A Survey of Empirical Applications in Agricultural Economics", *Agricultural Economics*, 7, 185-208.

Battese, G.E. (1997), "A Note on the Estimation of Cobb-Douglas Production Functions When Some Explanatory Variables Have Zero Values", *Journal of Agricultural Economics*, 48, 250-252.

Battese, G.E., and T.J. Coelli (1988), "Prediction of Firm-Level Technical Efficiencies With a Generalised Frontier Production Function and Panel Data", *Journal of Econometrics*, 38, 387-399.

Battese, G.E., and T.J. Coelli (1992), "Frontier Production Functions, Technical Efficiency and Panel Data: With Application to Paddy Farmers in India", *Journal of Productivity Analysis*, 3, 153-169.

Battese, G.E., and T.J. Coelli (1993), "A Stochastic Frontier Production Function Incorporating a Model for Technical Inefficiency Effects", *Working Papers in Econometrics and Applied Statistics*, No. 69, Department of Econometrics, University of New England, Armidale.

Battese, G.E., and T.J. Coelli (1995), "A Model for Technical Inefficiency Effects in a Stochastic Frontier Production Function for Panel Data", *Empirical Economics,* 20, 325-332.

Battese, G.E., T.J. Coelli and T.C. Colby (1989), "Estimation of Frontier Production Functions and the Efficiencies of Indian Farms Using Panel Data From ICRISAT's Village Level Studies", *Journal of Quantitative Economics*, 5, 327-348.

Battese, G.E., and G.S. Corra (1977), "Estimation of a Production Frontier Model: With Application to the Pastoral Zone of Eastern Australia", *Australian Journal of Agricultural Economics*, 21, 169-179.

Battese, G.E., A.N. Rambaldi and G.H. Wan (1997), "A Stochastic Frontier Production Function with Flexible Risk Properties", *Journal of Productivity Analysis*, 8, 269-280.

Bauer, P.W. (1990), "Recent Developments in the Econometric Estimation of Frontiers", *Journal of Econometrics*, 46, 39-56.

Beattie, B.R., and C.R. Taylor (1985), *The Economics of Production*, Wiley, New York.

Berger, A.N., and L.J. Mester (1997) "Inside the Black Box: What Explains Differences in the Efficiencies of Financial Institutions?", *Journal of Banking and Finance*, 21, 895-947.

Bessent, A.M., and E.W. Bessent (1980), "Comparing the Comparative Efficiency of Schools through Data Envelopment Analysis", *Educational Administration Quarterly*, 16, 57-75.

Bjurek, H. (1996), "The Malmquist Total Factor Productivity Index", *Scandinavian Journal of Economics*, 98, 303-313.

Blackorby, C., and R. Russell (1979), "Will the Real Elasticity of Substitution Please Stand Up?", *American Economic Review*, 79, 882-888.

Boles, J.N. (1966), "Efficiency Squared - Efficiency Computation of Efficiency Indexes", *Proceedings of the 39th Annual Meeting of the Western Farm Economics Association*, pp. 137-142.

Bortkiewicz, L. von (1923), "zweck und Struktur einer Preisindexzahl", Nordisk Statistisk Tidskrift, 369-408.

Bratley, P., Fox, B.L. and L.E. Schrage (1983), *A Guide to Simulation*, Springer-Verlag, New York.

Breusch, T., and A. Pagan (1980), "The LM Test and Its Applications to Model Specification in Econometrics", *Review of Economic Studies*, 47, 239-254.

Call, S.T., and W.L. Holahan (1983), *Microeconomics*, 2nd Ed., Wadsworth, Belmont.

Caudill, S.B., and J.M. Ford (1993), "Biases in Frontier Estimation Due to Heteroskedasticity", *Economics Letters*, 41, 17-20.

Caudill, S.B., J.M. Ford and D.M. Gropper (1995), "Frontier Estimation and Firm-Specific Inefficiency Measures", *Journal of Business and Economic Statistics*, 13:1, 105-11.

Caves, D.W., L.R. Christensen and W.E. Diewert (1982a), "Multilateral Comparisons of Output, Input and Productivity Using Superlative Index Numbers", *Economic Journal*, 92, 73-86.

Caves, D.W., L.R. Christensen and W.E. Diewert (1982b), "The Economic Theory of Index Numbers and the Measurement of Input, Output and Productivity", *Econometrica*, 50, 1393-1414.

Chambers, R.G. (1988), *Applied Production Analysis: A Dual Approach*, Cambridge University Press, New York.

Chambers, R.G., Y. Chung and R. Fare (1996), "Benefit and Distance Functions", *Journal of Economic Theory*, 70, 407-419.

Chambers, R.G., and J. Quiggin (2000), *Uncertainty, Production, Choice and Agency: The State-Contingent Approach*, Cambridge University Press, Cambridge, UK.

Charnes, A., W.W. Cooper, B. Golany, L. Seiford and J. Stutz (1985), "Foundations of Data Envelopment Analysis for Pareto-Koopmans Efficient Empirical Production Functions", *Journal of Econometrics*, 30, 91-107.

Charnes, A., W.W. Cooper, A.Y. Lewin and L.M. Seiford (1995), *Data Envelopment Analysis: Theory, Methodology and Applications*, Kluwer Academic Publishers, Boston.

Charnes, A., W.W. Cooper and E. Rhodes (1978), "Measuring the Efficiency of Decision Making Units", *European Journal of Operational Research*, 2, 429-444.

Charnes, A., W.W. Cooper and E. Rhodes (1981), "Evaluating Program and Managerial Efficiency: An Application of Data Envelopment Analysis to Program Follow Through", *Management Science*, 27, 668-697.

Charnes, A., W.W. Cooper, J. Rousseau and J. Semple (1987), "Data Envelopment Analysis and Axiomatic Notions of Efficiency and Reference Sets", *Research Report CCS 558*, Center for Cybernetic Studies, The University of Texas at Austin, Austin.

Charnes, A., C.T. Clark, W.W. Cooper and B. Golany (1985), "A Developmental Study of Data Envelopment Analysis in Measuring the Efficiency of Maintenance Units in the U.S. Air Forces", In R.G. Thompson and R.M. Thrall (Eds.), *Annals of Operations Research*, 2, pp.95-112.

Chiang, A.C. (1984), *Fundamental Methods of Mathematical Economics*, 3rd edition, McGraw-Hill, Singapore.

Christensen, L.R., and W.H. Greene (1976), "Economies of Scale in US Electric Power Generation", *Journal of Political Economy*, 84, 655-676.

Christensen, L.R., and D.W. Jorgensen, (1969), "The Measurement of US Real Capital Input, 1929-1967', *Review of Income and Wealth*, Volume 15, Issue 4, 293-320.

Coelli, T.J. (1992), "A Computer Program for Frontier Production Function Estimation: FRONTIER, Version 2.0", *Economics Letters*, 39, 29-32.

Coelli, T.J. (1995), "Estimators and Hypothesis Tests for a Stochastic Frontier Function: A Monte Carlo Analysis", *Journal of Productivity Analysis*, 6, 247-268.

Coelli, T.J. (1996a), "A Guide to FRONTIER Version 4.1: A Computer Program for Frontier Production Function Estimation", *CEPA Working Paper 96/07*, Department of Econometrics, University of New England, Armidale.

Coelli, T.J. (1996b), "A Guide to DEAP Version 2.1: A Data Envelopment Analysis (Computer) Program", *CEPA Working Paper 96/08*, Department of Econometrics, University of New England, Armidale.

Coelli, T.J. (1996c), "Measurement of Total factor Productivity Growth and Biases in Technological Change in Western Australian Agriculture", *Journal of Applied Econometrics*, 11, 77-91.

Coelli, T.J. (1996d), "Assessing the Performance of Australian Universities using Data Envelopment Analysis", mimeo, Centre for Efficiency and Productivity Analysis, University of New England.

Coelli, T.J. (1998), "A Multi-stage Methodology for the Solution of Orientated DEA Models", *Operations Research Letters*, 23, 143-149.

Coelli, T.J. (2000), "On the Econometric Estimation of the Distance Function Representation of a Production Technology", Discussion Paper 2000/42, Center for Operations Research and Econometrics, Universite Catholique de Louvain.

Coelli, T.J. (2002), "A Comparison of Alternative Productivity Growth Measures: With Application to Electricity Generation", in Fox, K. (ed), *Efficiency in the Public Sector*, Kluwer Academic Publishers, Boston.

Coelli, T.J., A. Estache, S. Perelman and Lourdes Trujillo (2003), *A Primer on Efficiency Measurement for Utilities and Transport Regulators*, World Bank Institute, Washington D.C.

Coelli, T.J., and G.E. Battese (1996), "Identification of Factors which Influence the Technical Efficiency of Indian Farmers", *Australian Journal of Agricultural Economics*, 40(2), 19-44.

Coelli, T.J., and S. Perelman (1996), "Efficiency Measurement, Multiple-output Technologies and Distance Functions: With Application to European Railways", *CREPP Discussion Paper* no. 96/05, University of Liege, Liege.

Coelli, T.J., and S. Perelman (1999), "A Comparison of Parametric and Non-parametric Distance Functions: With Application to European Railways", *European Journal of Operational Research*, 117:326-339.

Coelli, T.J., and S. Perelman (2000), "Technical Efficiency of European Railways: A Distance Function Approach" *Applied Economics*, 32, 1967-1976.

Coelli, T.J., S. Perelman and E. Romano (1999), "Accounting for Environmental Influences in Stochastic Frontier Models: With Application to International Airlines", *Journal of Productivity Analysis*, 11, 251-273.

Coelli, T.J., and D.S.P. Rao (2001), "Implicit Value Shares in Malmquist TFP Index Numbers", *CEPA Working Papers*, No. 4/2001, School of Economics, University of New England, Armidale, pp. 27.

Coelli, T.J., and D.S.P. Rao (2005), "Total Factor Productivity Growth in Agriculture: A Malmquist Index Analysis of 93 Countries, 1980-2000", *Agricultural Economics*, 32, 115-134.

Cooper W.W., L.M. Seiford and K. Tone (2000), *Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software*, Kluwer Academic Publishers, Boston.

Cornes, R. (1992), *Duality and Modern Economics*, Cambridge University Press, Melbourne.

Cornwell, C., P. Schmidt and R.C. Sickles (1990), "Production Frontiers With Cross-sectional and Time-series Variation in Efficiency Levels", *Journal of Econometrics*, 46, 185-200.

Cowing, T.G., and V.K. Smith (1980), "The Estimation of a Production Technology: A Survey of Econometric Analyses of Steam-Electric Generation", *Land Economics*, 54, 156-186.

Cuesta, R.A. (2000), "A Production Model With Firm-Specific Temporal Variation in Technical Inefficiency: With Application to Spanish Dairy Farms", *Journal of Productivity Analysis*, 13, 139-158.

Debreu, G. (1951), "The Coefficient of Resource Utilisation", *Econometrica*, 19, 273-292.

Deller, S.C., D. I. Chicoine and N. Walzer (1988), "Economies of Size and Scope in Rural Low-Volume Roads" *The Review of Economics and Statistics* 70(3):459-465.

Denny, M., M. Fuss and L. Waverman (1981), "The Measurement and Interpretation of Total Factor Productivity in Regulated Industries with an Application to Canadian Telecommunications." In T.G. Cowing and R.E. Stevenson (eds.), *Productivity Measurement in Regulated Industries*, Academic Press New York, 179–218.

Deprins, D., L. Simar and H. Tulkens (1984), "Measuring Labour-Efficiency in Post Offices" in M. Marchand, P. Pestieau and H. Tulkens (Eds.), *The Performance of Public Enterprises: Concepts and Measurements*, North-Holland, Amsterdam.

Diewert, W.E. (1976), "Exact and Superlative Index Numbers", *Journal of Econometrics,* 4, 115-45.

Diewert, W.E. (1978), "Superlative Index Numbers and Consistency in Aggregation", *Econometrica,* 46, 883-900.

Diewert, W.E. (1980) "Aggregation Problems in the Measurement of Capital", In Usher, D. (Ed.), *The Measurement of Capital*, National Bureau of Economic Research, Chicago, 433-528.

Diewert, W.E. (1981), "The Economic Theory of Index Numbers: A Survey", In Deaton, A. (Ed.), *Essays in the Theory and Measurement of Consumer Behaviour (in Honour of Richard Stone)*, Cambridge University Press, New York, 163-208.

Diewert, W.E. (1983), "The Theory of the Output Price Index and the Measurement of Real Output Change", In W.E. Diewert and C. Montmarquette (Eds.), *Price Level Measurement,* Statistics Canada, 1039-1113.

Diewert, W.E. (1990), *Price Level Measurement,* North-Holland, Amsterdam.

Diewert, W.E. (1992), "Fisher Ideal Output, Input and Productivity Indexes Revisited", *Journal of Productivity Analysis*, 3, 211-248.

Diewert, W.E. (1998), "Index Number Theory Using Differences Instead of Ratios", Paper presented at Yale University Conference Honouring Irving Fisher, 8, May, 1998, Discussion *Paper No. 98-10*, Department of Economics, University of British Columbia, Vancouver, Canada.

Diewert, W.E. (2000), "Productivity Measurement using Differences rather than Ratios: A Note", *Discussion Paper No. 2000/1*, School of Economics, University of New South Wales, Sydney, Australia.

Diewert, W.E., and D.A. Lawrence (1999), "Progress in Measuring the Price and Quantity of Capital " June 1999, pp51. University of British Columbia Department of Economics *Discussion Paper 99/17*, Vancouver, Canada.

Diewert, W.E., and A.O. Nakamura (1993), *Essays in Index Number Theory, Volume 1*, Contributions to Economic Analysis Series, No. 217, North-Holland, Amsterdam.